(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平10－133828

(43)公開日　平成10年(1998)５月22日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　3/06	５４０	Ｇ０６Ｆ　3/06	５４０
	３０５		３０５Ｃ

審査請求　未請求　請求項の数10　ＯＬ　（全 31 頁）

(21)出願番号　特願平8－288335

(22)出願日　平成８年(1996)10月30日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(72)発明者　内堀　郁夫
東京都府中市東芝町１番地　株式会社東芝府中工場内

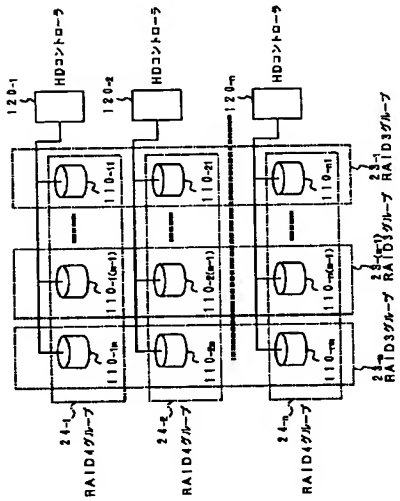
(72)発明者　金井　達徳
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

(72)発明者　島内　芳郎
東京都府中市東芝町１番地　株式会社東芝府中工場内

(74)代理人　弁理士　鈴江　武彦　（外６名）

(54)【発明の名称】　マルチメディアサーバ用ディスクアレイ装置

(57)【要約】
【課題】同一ディスク装置上で複数のディスクアレイ方式を可能とすると共に、２重のディスクアレイ保護を可能とし、信頼性の向上を図る。
【解決手段】ＨＤコントローラ１２０-1，…，１２０-nに接続される横方向配列ＨＤＤ１１０-11 ～１１０-1m，…，１１０-n1 ～１１０-nmによりＲＡＩＤ４グループ２４-1，…，２４-nを、縦方向の配列ＨＤＤ１１０-11 ～１１０-n1，…，１１０-1m ～１１０-nmによりＲＡＩＤ３グループ２３-1，…，２３-mをそれぞれ構成し、各ＨＤＤ１１０-11 ～１１０-nm の領域を、長期間繰り返し利用される大容量データの格納用のＲＡＩＤ３＆４領域と、一定期間のみ利用される大容量データの格納用のＲＡＩＤ３領域と、小容量データの格納用のＲＡＩＤ４領域とに分割し、ＲＡＩＤ３＆４領域をＲＡＩＤ３及びＲＡＩＤ４で、ＲＡＩＤ３領域をＲＡＩＤ３で、ＲＡＩＤ４領域をＲＡＩＤ４でそれぞれ保護する。



【特許請求の範囲】
【請求項１】　第１の方向及び第２の方向で表される２次元状に論理的に配置されたディスク装置の群であって、前記各ディスク装置のディスク領域が、第１の大容量データを格納するための第１の領域、前記第１の大容量データとは用途が異なる第２の大容量データを格納するための第２の領域、及び小容量データを格納するための第３の領域に分割して使用され、前記第１の方向のディスク装置の配列毎に、その配列内の全ての前記ディスク装置が並列にアクセスされる第１のディスクアレイグループが構成されると共に、前記第２の方向のディスク装置の配列毎に、その配列内の選択された前記ディスク装置がアクセスされる第２のディスクアレイグループが構成されるディスク装置の群と、
前記第２の方向（横方向）の前記ディスク装置の配列毎に設けられ、対応する配列内の前記各ディスク装置をアクセス制御するディスクコントローラと、
ホスト装置からの要求に従い前記各ディスクコントローラを制御する制御手段と、
前記制御手段の制御により前記各ディスクコントローラとの間のデータ入出力を行う入出力手段であって、前記各ディスクコントローラから読み出されたデータに基づくエラー修正と、前記ホスト装置から与えられる書き込みデータのエラー修正情報である第１のタイプのパリティの生成が可能な入出力手段とを具備し、
前記第１のディスクアレイグループ内の前記各ディスク装置の前記第１の領域には前記第１の大容量データの分割データまたはそのエラー修正情報である第１のタイプのパリティが配置され、
前記第１のディスクアレイグループ内の前記各ディスク装置の前記第２の領域には前記第２の大容量データの分割データまたはそのエラー修正情報である第１のタイプのパリティが配置され、
前記第２のディスクアレイグループ内の前記各ディスク装置の前記第３の領域には前記小容量データの分割データまたはそのエラー修正情報である第２のタイプのパリティが配置され、
前記第２のディスクアレイグループ内の前記ディスク装置の前記第１の領域には、同一グループ内の他の前記各ディスク装置の前記第１の領域に配置された分割データのエラー修正情報である第２のタイプのパリティが配置されることを特徴とするマルチメディアサーバ用ディスクアレイ装置。
【請求項２】　前記制御手段は、前記ホスト装置により前記第１の領域からのデータ読み出しが要求された場合、前記各ディスクコントローラに対して、目的とする前記第１のディスクアレイグループ内の前記各ディスク装置の前記第１の領域からのデータ読み出しを行わせると共に、読み出しエラー時には、該当するディスク装置が属する前記第２のディスクアレイグループ内の他の前記各ディスク装置の対応する前記第１の領域からデータを読み出してそのデータに基づくエラー修正を行わせる第１の方式によるリカバリ有りモードの読み出し制御を行い、エラー修正不可のディスクコントローラが１台だけ存在する場合には、他の前記各ディスクコントローラから読み出されたデータに基づくエラー修正を前記入出力手段に行わせるように構成されていることを特徴とする請求項１記載のマルチメディアサーバ用ディスクアレイ装置。
【請求項３】　前記制御手段は、前記ホスト装置により前記第１の領域からのデータ読み出しが要求された場合で且つ品質よりコンスタンス性を重視するとき、または前記ホスト装置により前記第２の領域からのデータ読み出しが要求された場合は、前記各ディスクコントローラに対して、目的とする前記第１のディスクアレイグループ内の前記各ディスク装置の前記第１の領域または前記第２の領域からのデータ読み出しのみを行わせ、読み出しエラーとなったディスクコントローラが１台だけ存在する場合には、他の前記各ディスクコントローラから読み出されたデータに基づくエラー修正を前記入出力手段に行わせるように構成されていることを特徴とする請求項２記載のマルチメディアサーバ用ディスクアレイ装置。
【請求項４】　前記制御手段は、前記ホスト装置により前記第１の領域へのデータ書き込みが要求された場合、前記ホスト装置から与えられる書き込みデータの前記第１のタイプのパリティを前記入出力手段により生成させると共に、前記書き込みデータの分割データまたは前記入出力手段により生成された第１のタイプのパリティを前記各ディスクコントローラにより目的とする前記第１のディスクアレイグループ内の前記各ディスク装置の前記第１の領域に書き込ませると共に、当該各ディスク装置がそれぞれ属する前記各第２のディスクアレイグループ内の前記各ディスク装置の対応する前記第１の領域のデータに基づく前記第２のタイプのパリティを生成させて対応するディスク装置の第１の領域に書き込ませることを特徴とする請求項１記載のマルチメディアサーバ用ディスクアレイ装置。
【請求項５】　前記第１の領域及び前記第２の領域には、前記ディスク装置のディスク領域において、前記第３の領域より外周側の領域が割り当てられることを特徴とする請求項１記載のマルチメディアサーバ用ディスクアレイ装置。
【請求項６】　前記制御手段は、前記ホスト装置により前記第３の領域からのデータ読み出しが要求された場合、前記ホスト装置により指定されたディスク装置に対応する前記ディスクコントローラに対し、当該ディスク装置の前記第３の領域からデータを読み出すための第２の方式によるリカバリ無しモードの読み出しを行わせ、読み出しエラーが発生したときには、前記指定されたデ

ィスク装置が属する前記第２のディスクアレイグループ内の他の前記各ディスク装置の対応する前記第３の領域からデータを読み出してそのデータに基づくエラー修正を行うための第２の方式によるリカバリモードの読み出しを行わせ、読み出しエラーのためにエラー修正不可のときには、前記指定されたディスク装置が属する前記第２のディスクアレイグループ内の前記指定されたディスク装置からのデータ読み出しと、読み出しエラー時に前記指定されたディスク装置が属する前記第２のディスクアレイグループ内の他の前記各ディスク装置の対応する前記第３の領域からデータを読み出してそのデータに基づくエラー修正とを行うための第２の方式によるリカバリ有りモードの読み出しを行わせるようにスケジュールすることを特徴とする請求項１記載のマルチメディアサーバ用ディスクアレイ装置。

【請求項７】　前記制御手段は、前記ホスト装置により前記第１の領域からのデータ読み出しが要求された場合、前記各ディスクコントローラに対して、目的とする前記第１のディスクアレイグループ内の前記各ディスク装置の前記第１の領域からのデータ読み出しのみ行わせる第２の方式によるリカバリ無しモードの読み出し制御を行い、読み出しエラーとなったディスクコントローラが１台だけ存在する場合には、他の前記各ディスクコントローラから読み出されたデータに基づくエラー修正を前記入出力手段に行わせ、読み出しエラーとなったディスクコントローラが２台以上存在する場合には、その読み出しエラーとなった各ディスクコントローラに対して、対応する前記第２のディスクアレイグループ内の該当する前記ディスク装置の前記第１の領域からのデータ読み出しを行わせると共に、読み出しエラー時には、当該第２のディスクアレイグループ内の他の前記各ディスク装置の対応する前記第１の領域からデータを読み出してそのデータに基づくエラー修正を行わせる第２の方式によるリカバリ有りモードの読み出し制御を行い、エラー修正不可のディスクコントローラが１台だけ存在する場合には、他の前記各ディスクコントローラから読み出されたデータに基づくエラー修正を前記入出力手段に行わせるように構成されていることを特徴とする請求項１記載のマルチメディアサーバ用ディスクアレイ装置。

【請求項８】　前記制御手段は、前記第１の領域への入出力要求と前記第２の領域への入出力要求との待ち行列である第１のキューを前記第１のディスクアレイグループ毎に、前記第１の領域に対する前記第２の方式によるリカバリ有りモードの読み出し要求と、前記第２のディスクアレイグループ内の前記各ディスク装置の前記第１の領域のデータに基づく前記第２のタイプのパリティを生成させて対応するディスク装置の第１の領域に書き込ませるための第２の方式によるパリティ生成・書き込み要求との待ち行列である第２のキューを前記第２のディスクアレイグループ毎に、それぞれ有し、

前記ホスト装置により前記第１の領域へのデータ書き込みが要求された場合には、前記ホスト装置から与えられる書き込みデータの前記第１のタイプのパリティを前記入出力手段により生成させると共に、前記書き込みデータの分割データまたは前記入出力手段により生成された前記第１のタイプのパリティを前記各ディスクコントローラにより目的とする前記第１のディスクアレイグループ内の前記各ディスク装置の前記第１の領域に書き込ませるための入出力要求を当該第１のディスクアレイグループに対応する前記第１のキューにつなぎ、この第１のキューにつないだ前記入出力要求の示す書き込みが正常終了したならば、前記第１のディスクアレイグループ内の前記各ディスク装置がそれぞれ属する前記各第２のディスクアレイグループ毎に、当該第２のディスクアレイグループを対象とする前記第２の方式によるパリティ生成・書き込み要求を、当該第２のディスクアレイグループに対応する前記第２のキューにつなぎ、前記ホスト装置に正常終了を通知するように構成されていることを特徴とする請求項７記載のマルチメディアサーバ用ディスクアレイ装置。

【請求項９】　前記制御手段は、前記第２の方式によるパリティ生成・書き込み要求を、同一の前記第２のキューの中で、どの前記第２の方式によるリカバリ有りモードの読み出し要求よりも先に実行される側につなぐことを特徴とする請求項８記載のマルチメディアサーバ用ディスクアレイ装置。

【請求項１０】　前記制御手段は、前記第３の領域への入出力要求の待ち行列である第３のキューを前記第２のディスクアレイグループ毎に更に有すると共に、前記第１のキューにつながれる要求の実行に要する時間に余裕値を加えた予め定められた第１の時間を記憶しておくための第１の時間記憶手段と、前記第１の時間、及び前記第２のキュー並びに前記第３のキューにつながれる各要求の実行に要する時間の上限値より大きい第２の時間を記憶しておくための第２の時間記憶手段と、前記第１の時間と実際の要求実行に要する時間の差の累積値である第３の時間を記憶しておくための第３の時間記憶手段とを更に有しており、

前記第１のキューにつなごうとする前記第１の領域または前記第２の領域への入出力要求の示す転送長が予め定められた一定長以上の場合には、当該入出力要求を前記一定長以下の転送長の複数の入出力要求に分割して前記第１のキューにつなぎ、前記第２のキュー及び前記第３のキューにつなぐ要求中には、その要求を他の要求との干渉がない状態で実行するのに要する時間の上限値を設定し、

通常は、前記第１のキューにつながれている要求を最優先として前記第１のディスクアレイグループが重複しないようにスケジュールすると共に、前記第２のキュー及び前記第３のキューにつながれている要求については、

その要求中に設定されている前記上限値が前記第１の時間より小さいものを対象にスケジュールし、少なくとも１つスケジュールできたならば、前記第１の時間を前記第３の時間に加えた値を新たな第３の時間として前記第３の時間記憶手段に記憶すると共に、そのスケジュールした要求の実行を制御し、その実行が全て終了するまでの経過時間を前記第３の時間から差し引いた値を新たな前記第３の時間として前記第３の時間記憶手段に記憶し、１つもスケジュールできなかったならば、前記第２の時間を新たな前記第３の時間として前記第３の時間記憶手段に記憶し、前記第３の時間が前記第２の時間以上となった場合には、前記第２のキュー及び前記第３のキューにつながれている要求のみを対象にスケジュールし、少なくとも１つスケジュールできたならば、そのスケジュールした要求の実行を制御して、その実行が全て終了した後に、１つもスケジュールできなかったならば直ちに、前記第３の時間記憶手段に初期値０を記憶するように構成されていることを特徴とする請求項９記載のマルチメディアサーバ用ディスクアレイ装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、複数のディスクドライブ（ディスク装置）を備えたマルチメディアサーバ用ディスクアレイ装置に関する。

【０００２】

【従来の技術】近年、複数のディスクドライブを備え、これら各ディスクドライブ（に装着されている記憶媒体）にデータを分散して格納することで、並列アクセスを可能としてアクセスの高速化を図ったディスクアレイ装置が開発されている。この種のディスクアレイ装置の代表的なものとしてＲＡＩＤ（Redundant Arrays of Inexpensive Disk）と称されるアーキテクチャを適用したディスクアレイ装置、即ちＲＡＩＤ装置が知られている。このＲＡＩＤアーキテクチャは、次のような文献、「D.Patterson,G.Gibson,and R.Katz,"A Case for Redundant Arrays of Inexpensive Disk(RAID),"ACM SIGMOD conference proceedings,Chicago,IL.,June 1-3,1988」及び「Randy H.Katz,Garth A.Gibson,and David A.Patterson,"DiskSystem Architectures for High Performance Computing,"Report No.UCB/CSD 89/497 March 1989」に記載されている。

【０００３】ＲＡＩＤアーキテクチャは、上記したようにデータを複数のディスクドライブに分散して格納することでアクセスの高速化を図る他、その分散格納されるデータに対応してパリティと称されるエラー修正情報（冗長データ）を格納することで、いずれかのディスクドライブに障害が発生したときに、パリティと残りの正常なディスクドライブのデータとから障害ディスクドライブのデータを回復（リカバリ）して信頼性の向上を図るようにしたものである。

【０００４】ところで、上記したようなディスクアレイ装置の利用形態の１つに、マルチメディア情報を格納する装置（マルチメディアサーバ）としての使用がある。このマルチメディア情報を格納する装置には、次のような３つの要求がある。

（１）信頼性の向上

まず、ビデオデータのように大量のデータを格納するためには、従来から知られているＲＡＩＤ装置（ディスクアレイ装置）以上に、冗長度が小さく、且つ信頼性が高い装置が要求されている。

【０００５】例えば、ディスクドライブの障害が発生した場合、ＲＡＩＤで保護されていない場合には、元のデータを再投入する必要が生じる。ところがビデオデータなどは、大量データであるため、ストライピング（データストリームを一定の単位で分割して分散配置すること）が行われることが多く、障害が発生したディスクドライブのデータだけでなく、関係する全てのデータの再投入を行う必要が生じる。この場合には、データのリカバリのために膨大な時間が必要となる。

【０００６】そのため、ＲＡＩＤの保護範囲を広げて一層の信頼性の向上を図ることで、このような事態の発生を防止することが要求される。

（２）異種データの混在

ＲＡＩＤアーキテクチャ（ＲＡＩＤ機構）は主にＲＡＩＤ０～５（ＲＡＩＤレベル０～５）の６つ（のレベル）に分類される。この６つのレベルのうち、ＲＡＩＤ３～５について簡単に述べる。

【０００７】まずＲＡＩＤ３は、比較的小さい単位（例えばバイト単位）でのストライピングを行うことで、１つのＩ／Ｏ要求に対して全ディスクドライブを並列（同時）にアクセスできるようにすると共に、ストライピングされたデータのパリティを専用のディスクドライブ（パリティ・ディスク）に格納する方式である。

【０００８】ＲＡＩＤ４は、比較的大きい単位（例えばディスクの物理セクタ以上の単位）でのストライピングを行うことで、複数のＩ／Ｏ要求の対象となるディスクドライブが要求毎に異なるようにして、各要求に対応するディスクドライブを独立にアクセスできるようにすると共に、ストライピングされたデータのパリティを専用のディスクドライブに格納する方式である。

【０００９】ＲＡＩＤ５は、ストライピングされたデータとパリティとがインターリーブされる点を除いてＲＡＩＤ４と同様である。以上のことから、ビデオデータのように、大量で、しかもディスクドライブの障害時にも一定の転送レートを保証しなければならないデータの格納には、ＲＡＩＤ３が適している。

【００１０】一方、静止画のように、比較的小量で、スループットが重視されるデータの格納には、ＲＡＩＤ４または５が適している。ところが、ビデオデータと静止画の両データを扱う場合には、ＲＡＩＤ３またはＲＡＩ

Ｄ４（５）の一方だけでは不具合が生じる。
【００１１】このため、同一ＲＡＩＤ装置（ディスクアレイ装置）上で、ＲＡＩＤ３とＲＡＩＤ４（５）の両機構（レベル）を実現することが要望される。
（３）ビデオデータに要求される、多様な、保全性と要求品質への対応
ビデオデータの利用形態にも、例えば、（ａ）監視システムなどでの応用のように一定時間のみ利用され、必要ならバックアップが取られた後、上書きされるようなデータと、（ｂ）長時間保存され、繰り返し利用されるデータとがあり、それぞれのデータで、耐障害性への要求が異なる。つまり、付加する冗長度、及び許されるオーバヘッドが異なる。
【００１２】また、ビデオデータの送出（転送）形態にも、例えば、（ａ）バッファメモリの少ないユーザ端末へ直接送出されるケースと、（ｂ）他のサーバ（ディスクアレイ装置）へ送られ（た後、そこからユーザ端末へ繰り返し送出され）るケースとがあり、それぞれのケースで、送出のための読み出し時の要求品質、及びコンスタント性が異なる。
【００１３】このため、同一ＲＡＩＤ装置（ディスクアレイ装置）上で、上記の多様な要求に応えることが要望される。そこで従来は、図２０或いは図２１に示すようなディスクアレイ装置が考えられている。
【００１４】図２０のディスクアレイ装置は、複数のディスクドライブ（ディスク装置）、例えば複数のハードディスクドライブ（ＨＤＤ）２０１を論理的に２次元状に配置して、コントローラ２０２の制御のもとで２次元のＲＡＩＤ方式を実現したものである。
【００１５】この図２０のディスクアレイ装置では、２重にＲＡＩＤ保護が図られることから、上記（１）の要求には応えられるが、上記（２），（３）の要求には応えられない。
【００１６】一方、図２１のディスクアレイ装置は、ＲＡＩＤ３で制御されるＨＤＤ２１１の列と、ＲＡＩＤ４，５で制御されるＨＤＤ２１２の列とを有し、ＲＡＩＤ３とＲＡＩＤ４，５の双方の機能を持つコントローラ２１３により、これらＨＤＤ２１１の列（ＲＡＩＤ３のグループ）とＨＤＤ２１２の列（ＲＡＩＤ４，５のグループ）を制御する構成としたものである。
【００１７】この図２２のディスクアレイ装置では、信頼性の点で改善されておらず、上記（１）の要求に応えていない。また、ディスクの負荷が均一にならず効率が悪く、ビデオデータなどの大量のデータと静止画などの小量のデータとの割合が変化するシステムではディスク内のスペースの使用効率が悪化する。つまり、上記（２）の要求に応えられない。また、上記（３）のような柔軟性（多様性）への要求にも応えられない。
【００１８】
【発明が解決しようとする課題】このように従来のディスクアレイ装置では、上記（１），（２），（３）で示したような要求に応えることができなかった。この発明は上記事情を考慮してなされたものでその目的は、ディスク装置の論理的な配列の方向によって異なる機能を持つ２次元のディスクアレイを実装することで、同一ディスク装置上で複数のディスクアレイ方式を可能とすると共に、２重のディスクアレイ保護を可能とし、信頼性の向上を図ると共に、異種データの混在を可能とし、しかも各データ種類・用途に応じた最適なアクセスが行えるマルチメディアサーバ用ディスクアレイ装置を提供することにある。
【００１９】
【課題を解決するための手段】本発明のディスクアレイ装置は、第１の方向及び第２の方向で表される２次元状に論理的に配置されたディスク装置の群であって、各ディスク装置のディスク領域が、第１の大容量データを格納するための第１の領域（以下、ＲＡＩＤ３＆４領域と称する）、上記第１の大容量データとは用途が異なる第２の大容量データ（以下、ＲＡＩＤ３領域と称する）を格納するための第２の領域、及び小容量データを格納するための第３の領域（以下、ＲＡＩＤ４領域と称する）に分割して使用され、上記第１の方向のディスク装置の配列（以下、第１の方向の配列と称する）毎に、その配列内の全てのディスク装置が並列にアクセスされる第１のディスクアレイグループ（以下、ＲＡＩＤ３グループと称する）が構成されると共に、上記第２の方向のディスク装置の配列（以下、第２の方向の配列と称する）毎に、その配列内の選択されたディスク装置がアクセスされる第２のディスクアレイグループ（ＲＡＩＤ４グループ）が構成されるディスク装置の群と、上記第２の方向の配列毎に設けられ、対応する配列内の各ディスク装置をアクセス制御するディスクコントローラと、ホスト装置からの要求に従い各ディスクコントローラを制御する制御手段と、この制御手段の制御により各ディスクコントローラとの間のデータ入出力を行う入出力手段（ＲＡＩＤ機構）であって、各ディスクコントローラから読み出されたデータに基づくエラー修正（以下、ＲＡＩＤ３によるリカバリと称する）と、ホスト装置から与えられる書き込みデータのエラー修正情報である第１のタイプのパリティ（以下、ＲＡＩＤ３のパリティと称する）が生成可能な入出力手段（ＲＡＩＤ機構）とを備えており、上記ＲＡＩＤ３グループ（第１のディスクアレイグループ）内の各ディスク装置のＲＡＩＤ３＆４領域には上記第１の大容量データの分割データまたはそのエラー修正情報であるＲＡＩＤ３のパリティ（第１のタイプのパリティ）が配置され、上記第１のディスクアレイグループ内の各ディスク装置のＲＡＩＤ３領域には上記第２の大容量データの分割データまたはそのエラー修正情報であるＲＡＩＤ３のパリティ（第１のタイプのパリティ）が配置され、上記ＲＡＩＤ４グループ（第２のディ

スクアレイグループ）内の各ディスク装置のＲＡＩＤ４領域には上記小容量データの分割データまたはそのエラー修正情報である第２のタイプのパリティ（以下、ＲＡＩＤ４のパリティと称する）が配置され、上記ＲＡＩＤ４グループ内のディスク装置のＲＡＩＤ３＆４領域には、同一グループ内の他の各ディスク装置のＲＡＩＤ３＆４領域に配置された分割データのエラー修正情報であるＲＡＩＤ４のパリティが配置されることを特徴とする。

【００２０】このようなディスクアレイ装置においては、第１の大容量データが格納されるＲＡＩＤ３＆４領域はＲＡＩＤ３とＲＡＩＤ４の両方で保護される。したがってＲＡＩＤ３＆４領域は、耐障害性への要求がより高い、例えば長期間繰り返し利用される通常のビデオデータの格納に適している。また、第２の大容量データが格納されるＲＡＩＤ３領域はＲＡＩＤ３で保護される。したがってＲＡＩＤ３領域は、上記第１の大容量データに比べれば耐障害性への要求は低いが、オーバヘッドがより少ないことが要求される、例えば監視システムでの応用のように一定期間のみ利用されるビデオデータの格納に適している。一方、スループットが重視される小容量データ、例えば静止画等のデータは、各ディスク装置のＲＡＩＤ４領域に格納することでＲＡＩＤ４が適用可能となるため、スループットの向上が図れる。

【００２１】また本発明は、上記ディスクアレイ装置において、第１の大容量データの読み出し時の耐障害性を実現するため、ホスト装置により上記ＲＡＩＤ３＆４領域からのデータ読み出しが要求された場合に、制御手段は、上記各ディスクコントローラに対して、目的とするＲＡＩＤ３グループ内の各ディスク装置のＲＡＩＤ３＆４領域からのデータ読み出しを行わせると共に、読み出しエラー時には、該当するディスク装置が属するＲＡＩＤ４グループ内の他の各ディスク装置の対応するＲＡＩＤ３＆４領域からデータを読み出してそのデータに基づくエラー修正を行わせる第１の方式によるリカバリ有りモードの読み出し（以下、ＲＡＩＤ４のリカバリ有りの読み出しと称する）制御を行い、エラー修正不可のディスクコントローラが１台だけ存在する場合には、他の前記各ディスクコントローラから読み出されたデータに基づくエラー修正（ＲＡＩＤ３によるリカバリ）を上記入出力手段に行わせることを特徴とする。

【００２２】このようなディスクアレイ装置においては、ＲＡＩＤ３による読み出しでエラーとなったディスク装置が複数存在したとしても、ＲＡＩＤ４による読み出しを行ってＲＡＩＤ４のパリティを用いたエラー修正を実行することでリカバリすることが可能となる。但し、ＲＡＩＤ３による読み出しでエラーとなったディスク装置が１つだけの場合には、上記入出力手段（ＲＡＩＤ機構）でのＲＡＩＤ３によるリカバリで対処可能なため、対応するディスクコントローラ１台のみが行うリカバリ処理に要する時間が無駄となる。

【００２３】そこで、このような無駄を無くすため、ホスト装置によりＲＡＩＤ３＆４領域からのデータ読み出しが要求された場合に、上記制御手段は、上記各ディスクコントローラに対して、目的とするＲＡＩＤ３グループ内の各ディスク装置のＲＡＩＤ３＆４領域からのデータ読み出しのみ行わせる第２２の方式によるリカバリ無しモードの読み出し（ＲＡＩＤ４によるリカバリ無しモードの読み出し）制御を行い、読み出しエラーとなったディスクコントローラが１台だけ存在する場合には、他の各ディスクコントローラから読み出されたデータに基づくエラー修正（ＲＡＩＤ３による復旧）を入出力手段（ＲＡＩＤ機構）に行わせ、読み出しエラーとなったディスクコントローラが２台以上存在する場合に、その読み出しエラーとなった各ディスクコントローラに対して、対応するＲＡＩＤ４グループ内の該当するディスク装置のＲＡＩＤ３＆４領域からのデータ読み出しを行わせると共に、読み出しエラー時には、当該ＲＡＩＤ４グループ内の他の各ディスク装置の対応するＲＡＩＤ３＆４領域からデータを読み出してそのデータに基づくエラー修正を行わせる第２の方式によるリカバリ有りモードの読み出し（ＲＡＩＤ４によるリカバリ有りモードの読み出し）制御を行い、エラー修正不可のディスクコントローラが１台だけ存在する場合には、他の各ディスクコントローラから読み出されたデータに基づくエラー修正を上記入出力手段に行わせることを特徴とする。

【００２４】また本発明は、ホスト装置によりＲＡＩＤ３＆４領域からのデータ読み出しが要求された場合でも、品質よりコンスタンス性を重視するとき、例えば読み出したデータを直接ユーザに送信するような場合には、他のサーバに送信するために読み出す場合と異なって、ＲＡＩＤ３領域からのデータ読み出しが要求された場合と同様に、ＲＡＩＤ３による読み出しのみを行う、即ち上記各ディスクコントローラに対して、目的とするＲＡＩＤ３グループ内の各ディスク装置の前記第１の領域からのデータ読み出しのみを行わせ、読み出しエラーとなったディスクコントローラが１台だけ存在する場合には、他の前記各ディスクコントローラから読み出されたデータに基づくエラー修正を上記入出力手段に行わせることを特徴とする。

【００２５】このように本発明においては、ＲＡＩＤ３とＲＡＩＤ４とで２重に保護されているＲＡＩＤ３＆４領域から、用途に応じて、２種類の読み出し方法のいずれかが選択可能である。

【００２６】また本発明は、ホスト装置によりＲＡＩＤ４領域からのデータ読み出しが要求された場合に、ホスト装置により指定されたディスク装置に対応するディスクコントローラに対し、当該ディスク装置のＲＡＩＤ４領域からのデータ読み出しのみ行わせ（ＲＡＩＤ４によるリカバリ無しモードの読み出しを行わせ）、読み出し

エラーが発生したときには、上記指定されたディスク装置が属するＲＡＩＤ４グループ内の他の各ディスク装置の対応するＲＡＩＤ４領域からのデータ読み出しとそのデータに基づくエラー修正を行わせ（ＲＡＩＤ４によるリカバリモードの読み出しを行わせ）、読み出しエラーのためにエラー修正不可のときには、上記指定されたディスク装置が属するＲＡＩＤ４グループ内の当該ディスク装置からのデータ読み出しと、読み出しエラー時に当該ディスク装置が属するＲＡＩＤ４グループ内の他の各ディスク装置の対応するＲＡＩＤ４領域からデータを読み出してそのデータに基づくエラー修正とを行わせる（ＲＡＩＤ４によるリカバリ有りモードの読み出しを行わせる）ようにスケジュールすることを特徴とする。

【００２７】このように、動作するディスク装置を細かく多段階に制御することで、大容量データの読み出しのスケジューリングに影響を与えないように、ＲＡＩＤ４領域からの小容量データの読み込みを行う機会を増加させることが可能となる。

【００２８】また本発明は、大容量データが格納されるＲＡＩＤ３＆４領域及びＲＡＩＤ３領域を、小容量データが格納されるＲＡＩＤ４領域より外周側のディスク領域に割り当てたことを特徴とする。

【００２９】このように、小容量データをディスクの内周部に、大容量データをそれより外周部に配置することにより、即ち内周部に比べてトラック当たりのデータ量が多い領域に大容量データを配置することにより、大容量データの配信の性能を向上させることができる。

【００３０】また本発明は、ホスト装置によりＲＡＩＤ３＆４領域へのデータ書き込みが要求された場合に、上記制御手段は、ホスト装置から与えられる書き込みデータの第１のタイプのパリティ（ＲＡＩＤ３のパリティ）を上記入出力手段により生成させると共に、当該書き込みデータの分割データまたは上記生成されたパリティを上記各ディスクコントローラにより目的とするＲＡＩＤ３グループ内の各ディスク装置のＲＡＩＤ３＆４領域に書き込ませると共に、当該各ディスク装置がそれぞれ属する各ＲＡＩＤ４グループ内の各ディスク装置の対応するＲＡＩＤ３＆４領域のデータに基づく第１のタイプのパリティ（ＲＡＩＤ４のパリティ）を生成させて対応するディスク装置のＲＡＩＤ３＆４領域に書き込ませることを特徴とする。

【００３１】このように本発明においては、ＲＡＩＤ３＆４領域に書き込まれるデータに対しては、ＲＡＩＤ３のパリティとＲＡＩＤ４のパリティの２種のパリティが生成され、ＲＡＩＤ３とＲＡＩＤ４とで２重に保護できる。

【００３２】また本発明は、上記したＲＡＩＤ３＆４領域へのデータ書き込みでは、ＲＡＩＤ４のパリティの生成・書き込みに時間を要することから、その不具合を解消するために、ＲＡＩＤ３＆４領域への入出力要求とＲＡＩＤ３領域への入出力要求との待ち行列である第１のキュー（キュー＃１）をＲＡＩＤ３グループ毎に、ＲＡＩＤ３＆４領域に対するＲＡＩＤ４によるリカバリ有りモードの読み出し要求と、ＲＡＩＤ４によるパリティ生成・書き込み要求との待ち行列である第２のキュー（キュー＃２）をＲＡＩＤ４グループ毎に、それぞれ上記制御手段に持たせ、ホスト装置によりＲＡＩＤ３＆４領域へのデータ書き込みが要求された場合には、ホスト装置から与えられる書き込みデータのＲＡＩＤ３のパリティを上記入出力手段により生成させると共に、当該書き込みデータの分割データまたは上記生成されたパリティを上記各ディスクコントローラにより目的とするＲＡＩＤ３グループ内の各ディスク装置のＲＡＩＤ３＆４領域に書き込ませるための入出力要求を当該ＲＡＩＤ３グループに対応する上記第１のキューにつなぎ、この第１のキューにつないだ入出力要求の示す書き込みが正常終了したならば、上記ＲＡＩＤ３グループ内の各ディスク装置がそれぞれ属する上記各ＲＡＩＤ４グループ毎に、当該ＲＡＩＤ４グループ内の各ディスク装置の対応するＲＡＩＤ３＆４領域のデータに基づくＲＡＩＤ４のパリティを生成させて対応するディスク装置のＲＡＩＤ３＆４領域に書き込ませるためのＲＡＩＤ４によるパリティ生成・書き込み要求を、当該ＲＡＩＤ４グループに対応する上記第２のキューにつないで上記ホスト装置に正常終了を通知し、上記ＲＡＩＤ４によるパリティ生成・書き込み要求の実行をホスト装置への正常終了通知より後にしたことを特徴とする。

【００３３】本発明においては、ＲＡＩＤ３＆４領域へのデータ書き込み時の見かけ上の性能向上が可能となる。なお、ＲＡＩＤ４によるパリティ生成・書き込み要求を実際に実行した結果、異常終了した場合には、ホスト装置からの要求とは非同期に制御手段からホスト装置に異常終了を通知すればよい。

【００３４】また、上記ＲＡＩＤ４によるパリティ生成・書き込み要求は、同一の第２のキューの中では、どのＲＡＩＤ４によるリカバリ有りモードの読み出し要求よりも先に実行される側につながれるようにすると、誤ったデータを読み込むことになるのを防ぐことができる。

【００３５】また本発明は、ＲＡＩＤ４領域への入出力要求の待ち行列である第３のキュー（キュー＃３）をＲＡＩＤ４アレイグループ毎に制御手段に更に持たせると共に、上記第１のキューにつながれる要求の実行に要する時間に余裕値を加えた予め定められた第１の時間（Ｔ）を記憶しておくための第１の時間記憶手段（Ｔレジスタ）と、上記第１の時間、及び上記第２のキュー並びに第３のキューにつながれる各要求の実行に要する時間の上限値より大きい第２の時間（Ａ）を記憶しておくための第２の時間記憶手段（Ａレジスタ）と、上記第１の時間と実際の要求実行に要する時間の差の累積値である第３の時間（ｔ）を記憶しておくための第３の時間記

憶手段（ｔレジスタ）とについても制御手段に持たせ、第１のキューにつなごうとするＲＡＩＤ３＆４領域またはＲＡＩＤ３領域への入出力要求の示す転送長が予め定められた一定長以上の場合には、当該入出力要求を上記一定長以下の転送長の複数の入出力要求に分割して第１のキューにつなぎ、第２のキュー及び第３のキューにつなぐ要求中には、その要求を他の要求との干渉がない状態で実行するのに要する時間の上限値を設定し、通常は、第１のキューにつながれている要求を最優先としてＲＡＩＤ３グループが重複しないようにスケジュールすると共に、第２のキュー及び第３のキューにつながれている要求については、その要求中に設定されている上限値が上記第１の時間（Ｔ）より小さいものを対象にスケジュールし、少なくとも１つスケジュールできたならば（ここでのスケジュールを第１のスケジュールと称する）、上記第１の時間（Ｔ）を上記第３の時間（ｔ）に加えた値を新たな第３の時間（ｔ）として上記第３の時間記憶手段に記憶すると共に、そのスケジュールした要求の実行を制御し、その実行が全て終了するまでの経過時間を上記第３の時間（ｔ）から差し引いた値を新たな第３の時間（ｔ）として上記第３の時間記憶手段に記憶し、１つもスケジュールできなかったならば、上記第２の時間（Ａ）を新たな第３の時間（ｔ）として上記第３の時間記憶手段に記憶し、上記第３の時間（ｔ）が上記第２の時間（Ａ）以上となった場合に限り、第２のキュー及び第３のキューにつながれている要求のみを対象にスケジュールし、少なくとも１つスケジュールできたならば（ここでのスケジュールを第２のスケジュールと称する）、そのスケジュールした要求の実行を制御して、その実行が全て終了した後に、１つもスケジュールできなかったたならば直ちに、上記第３の時間記憶手段に初期値０を記憶するようにしたことを特徴とする。

【００３６】本発明において、上記第１のスケジュールにより対応する要求が実行されると、その要求の実行の経過時間はＴ（第１の時間）より小さいことが保証されるため、その要求の実行後に行われるｔ（第３の時間）の更新の結果、当該ｔは増加する。したがって、このような処理が繰り返される毎に、上記ｔは増加していき、そのうちにｔ≧Ａとなる。すると、第２のキュー及び第３のキューにつながれている要求のみを対象とする第２のスケジュールが行われる。ここでは、その要求の実行に要する時間の上限値がＴより大きいもの（但し、Ａよりは小さい）もスケジュールの対象となる。この第２のスケジュールの結果、スケジューリングされた要求が実行されると、上記ｔは初期値０に更新されるため、今度は第１のスケジュールが行われることになり、第２のスケジュールが続けて行われることはない。

【００３７】したがって本発明においては、第１のキューが空きの状態で新しい要求が登録された場合には、その要求は（Ａ＋Ｔ）時間以内にスケジュールされる。また１つの第１のキューに例えばｐ個の要求がつながっている状態で、新しい要求が登録された場合には、先頭の要求がスケジュールされるまでの最長時間は（Ａ＋Ｔ）であり、ｐ個の要求の実行に要する時間の上限はｐＴであることから、新たに登録された要求は（Ａ＋（ｐ＋１）Ｔ）時間以内にスケジュールされる。即ち、第１のキューへの要求の頭出し時間が保証できる。

【００３８】また、本発明においては、第１のキューが空にならないようにホスト装置から制御することより、Ｔ時間に１回の割合で入出力要求を実行することができる。即ち、第１のキューへの入出力のレートが保証できる。但し、リカバリ有りモードの入出力要求を発行した場合で、第２のキューにつなぎかえられた場合には、その終了の通知は遅れる。

【００３９】また、本発明においては、第１のキューに要求がつながれている状態でも、第２及び第３のキューの要求にはスケジュールされる機会が訪れる。以上により、ビデオデータのような大容量データの入出力については、一定の入出力レートを保証し、且つ余裕があれば、静止画などの小容量データの入出力も並行して実行できる。

【００４０】

【発明の実施の形態】以下、本発明の実施の形態につき図面を参照して説明する。図１は本発明の一実施形態に係るマルチメディアサーバ用ディスクアレイ装置１０の構成を示すブロック図である。

【００４１】図１のディスクアレイ装置１０は、同一装置上で例えばＲＡＩＤ３とＲＡＩＤ４の両機構を実現するものであり、ディスクドライブ（ディスク装置）群、例えばハードディスクドライブ（ＨＤＤ）群１１と、ＨＤ（ハードディスク）コントローラ１２０-1～１２０-nからなるＨＤコントローラ部１２と、バッファ１３０-1～１３０-nからなるバッファ部１３と、ＲＡＩＤ機構１４と、制御部１５とから構成される。

【００４２】ＨＤＤ群１１は、論理的にｍ×ｎの２次元配列をなす、ｍ×ｎ個のＨＤＤ、即ちＨＤＤ１１０-11～１１０-1m，１１０-21～１１０-2m，…，１１０-n1～１１０-nmからなる。

【００４３】ここで、ＨＤＤ群１１におけるｘ方向（横方向）の配列であるＨＤＤ１１０-11～１１０-1m，１１０-21～１１０-2m，…，１１０-n1～１１０-nmは、ＨＤコントローラ１２０-1，１２０-2，…，１２０-nに接続されている。ＨＤＤ１１０-11～１１０-1m，…，１１０-n1～１１０-nm（に装着される図３に示す各ディスク３０の内周部ＩＡの領域、即ちＲＡＩＤ４領域３１３）は、図２に示すように、それぞれＲＡＩＤ４のグループ２４-1，…，２４-nを構成する。また、ＨＤＤ群１１におけるｙ方向（縦方向）の配列であるＨＤＤ１１０-11～１１０-n1，…，１１０-1m～１１０-nm（に装着される図３に示す各ディスク３０の外周部ＯＡ

の領域、即ちＲＡＩＤ３＆４領域３１１とＲＡＩＤ３領域３１２）は、図２に示すように、それぞれＲＡＩＤ３のグループ２３-1，…，２３-mを構成する。

【００４４】なお、本実施形態では、１つのＨＤコントローラ１２０-i（ｉ＝１～ｎ）下に１つＲＡＩＤ４グループ２４-iが接続された構成を適用しているが、全てのＨＤコントローラ１２０-i（ｉ＝１～ｎ）に同数の複数のグループが接続される構成であっても構わない。

【００４５】ＲＡＩＤ３グループ２３-mを構成するＨＤＤ１１０-1m ～１１０-nm （に装着される図３に示す各ディスク３０の内周部ＩＡの領域、即ちＲＡＩＤ４領域３１３と、外周部ＯＡの領域の一部をなすＲＡＩＤ３＆４領域３１１）は、ＲＡＩＤ４グループ２４-1～２４-nのパリティ格納用に用いられ、ＲＡＩＤ４グループ２４-nを構成するＨＤＤ１１０-n1 ～１１０-nm （に装着される図３に示す各ディスク３０の外周部ＯＡの領域、即ちＲＡＩＤ３＆４領域３１１とＲＡＩＤ３領域３１２）は、ＲＡＩＤ３グループ２３-1～２３-mのパリティ格納用に用いられる。

【００４６】ＨＤコントローラ部１２内のＨＤコントローラ１２０-1～１２０-nは、制御部１５からの指示により、ＲＡＩＤ４のコントローラとして、自身に接続されているＲＡＩＤ４グループ２４-1～２４-n内のＨＤＤを制御する。またＨＤコントローラ１２０-1～１２０-nは、制御部１５からの指示により、ＲＡＩＤ機能を持たないコントローラとしても動作する。

【００４７】バッファ１３内のバッファ１３０-1～バッファ１３０-nは、ＨＤコントローラ１２０-1～１２０-nにより入出力されるデータを一時的に格納するのに用いられる。

【００４８】ＲＡＩＤ機構１４は、制御部１５からの指示により、ＲＡＩＤ３の機能を実行することも、バッファ１３内のバッファ１３０-1～１３０-nの内容をそのまま入出力することも可能なように構成されている。

【００４９】制御部１５は、図示せぬホスト装置からの指示を受け、ＨＤコントローラ部１２及びＲＡＩＤ機構１４を通して装置全体を制御する。この制御部１５は、後述する境界情報等を記憶しておくための不揮発性メモリ１５０を内蔵している。

【００５０】さて本実施形態では、ＨＤＤ群１１内の各ＨＤＤ１１０-ij （ｉ＝１～ｎ，ｊ＝１～ｍ）の領域、更に具体的に述べるならばＨＤＤ１１０-ij に装着される図３に示すようなディスク３０の領域を、次の３種類に分けて使用している。

（ａ）ＲＡＩＤ３及び４の双方で保護される領域３１１

この領域（以下、ＲＡＩＤ３＆４領域と称する）３１１は、通常のビデオデータを格納するのに用いられる。この領域３１１への書き込みの際には、ＲＡＩＤ３及び４の双方のパリティが生成される。

【００５１】この領域３１１からの読み込みには、次の２つのモードがある。第１は、直接ユーザにビデオデータを送信するために読み出すような場合に、品質よりコンスタント性、システムのオーバヘッド軽減を目指して、ＲＡＩＤ４によるリカバリを行わないようにしたモード（以下、リカバリ無しモードと称する）である。

【００５２】第２は、他のサーバ（ディスクアレイ装置）にビデオデータを送信するために読み出すような場合に、品質を重視してＲＡＩＤ４によるリカバリを行うようにしたモード（以下、リカバリ有りモードと称する）である。

【００５３】いずれの場合にも、ＲＡＩＤ３によるリカバリは行われる。

（ｂ）ＲＡＩＤ３のみで保護される領域３１２

この領域（以下、ＲＡＩＤ３領域と称する）３１２は、監視システムへの応用のような、短時間保存されるビデオデータなどを格納するのに用いられる。この領域３１２への書き込みの際には、ＲＡＩＤ３のパリティのみが生成される。また、この領域３１２からの読み出しの際には、ＲＡＩＤ３によるリカバリが行われる。

（ｃ）ＲＡＩＤ４のみで保護される領域３１３

この領域（以下、ＲＡＩＤ４領域と称する）３１３は、静止画に代表される小量のデータを格納するのに用いられる。この領域３１３への書き込みの際には、ＲＡＩＤ４のパリティのみが生成される。また、この領域３１３からの読み出しの際には、ＲＡＩＤ４によるリカバリが行われる。

【００５４】本実施形態において、（各ＨＤＤ１１０-ij 内の）ディスク３０は、内周部ＩＡと、その外側の領域（内周部ＩＡを除く領域）である外周部ＯＡとに分けて管理される。内周部ＩＡはＲＡＩＤ４領域３１３に割り当てられ、静止画等の小量データを配置（格納）するのに用いられる。したがって、内周部ＩＡはＲＡＩＤ４で制御される。一方、外周部ＯＡはＲＡＩＤ３＆４領域３１１とＲＡＩＤ３領域３１２とに割り当てられ、ビデオデータ等の大量データを配置するのに用いられる。したがって、外周部ＯＡはＲＡＩＤ３或いはＲＡＩＤ３＆４で制御される。

【００５５】一般に、ディスク３０では、内周側より外周側の方がトラック当たりのデータ量が多いことから、外周側の方が読み出し速度（転送速度）が高速である。したがって、上記したように外周部ＯＡ側にビデオデータ等の大量データを配置する構成とすることで、ビデオデータ等の配信の性能を向上することができる。一方、静止画等の小量データは、上記したように内周部ＩＡ側に配置する構成とする。小量データは、（大量データに比べて）Ｉ／Ｏ時間（データの入出力に要する時間）の中に占めるシーク、回転待ち時間の割合が大きいため、転送時間が長くなっても性能低下の割合が少ない。

【００５６】以上により、システム全体の性能が向上する。なお、図３では、ＲＡＩＤ３領域３１２がＲＡＩＤ

３＆４領域３１１の外側となっているが、その逆であっても構わない。

【００５７】内周部ＩＡと外周部ＯＡとの境界の情報は、全てのＨＤＤ１１０-ij （ｉ＝１～ｎ，ｊ＝１～ｍ）内の全てのディスク３０に共通であり、図１中の制御部１５の有する不揮発性メモリ１５０内に確保された境界情報領域（以下、Ｂレジスタと称する）１５１に記憶される。この不揮発性メモリ１５０内のＢレジスタ１５１の境界情報は、ホスト装置からの命令で書き換えることが可能である。

【００５８】次に、本実施形態における動作を、（１）ＲＡＩＤ３＆４領域３１１からのリカバリ有りモードでの読み出し、（２）ＲＡＩＤ３＆４領域３１１への書き込み、（３）ＲＡＩＤ３領域３１２からの読み出し、またはＲＡＩＤ３＆４領域３１１からのリカバリ無しモードでの読み出し、（４）ＲＡＩＤ３領域３１２への書き込み、（５）ＲＡＩＤ４領域３１３からの読み出し、（６）ＲＡＩＤ４領域３１３への書き込みについて、順に説明する。

（１）ＲＡＩＤ３＆４領域３１１からのリカバリ有りモードでの読み出し

まず、ＲＡＩＤ３＆４領域３１１からのリカバリ有りモードでの読み出しについて、図４のフローチャートを参照して説明する。

【００５９】制御部１５は、ホスト装置からの指示が、通常のビデオデータ等、長期間保存されて繰り返し利用される大量データを他のサーバ（ディスクアレイ装置）に送信するためにＲＡＩＤ３＆４領域３１１からの読み出しを行うものである場合、ＨＤコントローラ部１２内の各ＨＤコントローラ１２０-1～１２０-nに対して、ＲＡＩＤ４によるリカバリ機能有りのリードのための命令（read命令）を発行する（ステップＳ１）。ここで、各ＨＤコントローラ１２０-i（ｉ＝１～ｎ）に対する命令（read命令）には、リードの対象となるＨＤＤ１１０-ij （ｊは１～ｍのいずれか）と当該ＨＤＤ１１０-ij 内のディスク領域の情報（領域先頭位置へのシークアドレスと、転送長の情報）が含まれている。

【００６０】各ＨＤコントローラ１２０-iは、制御部１５から与えられる命令（read命令）で要求されたＨＤＤ１１０-ij からのデータ読み出しを行い（ステップＳ２）、正常に読めたか否かをチェックする（ステップＳ３）。

【００６１】もし、エラーがなかったならば、ＨＤコントローラ１２０-iはステップＳ２で読み出したデータをバッファ部１３内のバッファ１３０-iに格納し、制御部１５に正常終了を通知する（ステップＳ４）。

【００６２】もし、エラーが有ったならば、ＨＤコントローラ１２０-iは、ＲＡＩＤ４によるリカバリを実行する（ステップＳ５）。即ちＨＤコントローラ１２０-iは、ＲＡＩＤ４グループ２４-i内のＨＤＤ１１０-ij を除く他のＨＤＤ１１０-i1 ～１１０-im を対象に、ステップＳ１で行ったのと同じ読み出しを行い、全て正常に読めたならば、リカバリ可能であるものと判断し、その読み出したデータの排他的論理和をとってＨＤＤ１１０-ij のデータをリカバリする。

【００６３】ＨＤコントローラ１２０-iは、ＲＡＩＤ４によるリカバリができた場合（ステップＳ６）、リカバリしたデータをバッファ部１３内のバッファ１３０-iに格納し、制御部１５に正常終了を通知する（ステップＳ７）。これに対し、１つでも正常に読めなかったＨＤＤが存在するならば、ＨＤコントローラ１２０-iはＲＡＩＤ４によるリカバリはできないものとして、制御部１５に対してリードエラー（リカバリ不可）を通知する（ステップＳ８）。

【００６４】制御部１５は、各ＨＤコントローラ１２０-iからのエラー通知（リカバリ不可通知）を監視しており（ステップＳ９）、エラーを通知したＨＤコントローラ１２０-iの数が０であるならば、バッファ部１３内のバッファ１３０-1～１３０-nには、ＨＤコントローラ１２０-1～１２０-nに要求したデータが格納されているものと判断してステップＳ１０に進む。このステップＳ１０では、制御部１５はＲＡＩＤ機構１４に対してＲＡＩＤ３の機能を起動させ、バッファ１３０-1～１３０-(n-1)のデータを合成させた後、その合成後のデータをホスト装置に出力する。

【００６５】また、エラーを通知したＨＤコントローラ１２０-iの数が１であるならば、制御部１５はステップＳ１１に進む。このステップＳ１１では、制御部１５はＲＡＩＤ機構１４に対してＲＡＩＤ３の機能を起動させ、エラーを通知した唯一のＨＤコントローラ１２０-iを除くＨＤコントローラ１２０-1～１２０-nに対応するバッファ１３０-1～１３０-nのデータから、エラーを通知したＨＤコントローラ１２０-iが読み出すべきデータをバッファ１３０-i内にリカバリさせ、バッファ１３０-1～１３０-(n-1)のデータを合成させた後、その合成後のデータをホスト装置に出力する。

【００６６】また、エラーを通知したＨＤコントローラ１２０-iの数が２以上であるならば、制御部１５はＲＡＩＤ機構１４のＲＡＩＤ３機能を起動してもリカバリ不可能であるとして、その旨をホスト装置に通知してエラー終了する（ステップＳ１２）。

（２）ＲＡＩＤ３＆４領域３１１への書き込み

次に、ＲＡＩＤ３＆４領域３１１への書き込みについて、図５のフローチャートを参照して説明する。

【００６７】制御部１５は、ホスト装置からの指示が、通常のビデオデータ等、長期間保存されて繰り返し利用される大量データのＲＡＩＤ３＆４領域３１１への書き込みを行うものである場合、まずＲＡＩＤ機構１４に対してＲＡＩＤ３の機能を起動させる。これによりＲＡＩＤ機構１４は、ホスト装置からのデータ（書き込みデー

タ）をＲＡＩＤ３でストライピングして、そのパリティを生成しながら、そのストライピングしたデータ（分割データ）及びパリティを、バッファ部１３内のそれぞれ対応するバッファ１３０-1～１３０-nに格納する（ステップＳ２１）。

【００６８】次に制御部１５は、HDコントローラ部１２内の各HDコントローラ１２０-1～１２０-nに対し、バッファ１３０-1～１３０-nに格納されたデータを、ＲＡＩＤ４でのパリティを生成しながら、ＲＡＩＤ３グループ２３-j（ｊは１～ｍ－１のいずれか）内の目的とするHDD１１０-1j ～１１０-nj に書き込むことを指示する（ステップＳ２２）。

【００６９】これを受けて各HDコントローラ１２０-i（ｉ＝１～ｎ）は、ＲＡＩＤ４でのパリティを生成しながら、バッファ１３０-iのデータをＲＡＩＤ３グループ２３-j内の目的とするHDD１１０-ij （即ちＲＡＩＤ４グループ２４-i内の目的とするHDD１１０-ij ）に書き込むと共に、生成したパリティをＲＡＩＤ４グループ２４-i内のHDD１１０-im に書き込む（ステップＳ２３）。ここで、ＲＡＩＤ４でのパリティの生成は次のように行われる。

【００７０】まずHDコントローラ１２０-iは、ＲＡＩＤ４グループ２４-i内のHDD１１０-il ～１１０-i(m-l) を対象に、目的HDD１１０-ij の書き込み対象領域に一致する領域からのデータ読み出しを行う。次にHDコントローラ１２０-iは、目的HDD１１０-ij から読み出したデータをバッファ１３０-iのデータで更新した後、この更新後のデータを含むHDD１１０-il ～１１０-i(m-l) からの読み出しデータの排他的論理和をとってパリティを生成する。そしてHDコントローラ１２０-iは、上記更新後のデータをＲＡＩＤ４グループ２４-i内の目的HD１１０-ij に書き込むと共に、生成したパリティをＲＡＩＤ４グループ２４-iのHDD１１０-im 内の領域（目的HDDの書き込み対象領域に一致する領域）に書き込む。

（３）ＲＡＩＤ３領域３１２からの読み出し、またはＲＡＩＤ３＆４領域３１１からのリカバリ無しモードでの読み出し

次に、ＲＡＩＤ３領域３１２からの読み出し、またはＲＡＩＤ３＆４領域３１１からのリカバリ無しモードでの読み出しについて、図６のフローチャートを参照して説明する。

【００７１】制御部１５は、ホスト装置からの指示が、監視システムで適用されるビデオデータ等、短期間保存される大量データをＲＡＩＤ３領域３１２から読み出す場合、或いは通常のビデオデータ等、長期間保存されて繰り返し利用される大量データを直接ユーザに送信するためにＲＡＩＤ３＆４領域３１１から読み出す場合、HDコントローラ部１２内の各HDコントローラ１２０-1～１２０-nに対して、ＲＡＩＤ４によるリカバリ機能無しのリードのための命令（read命令）を発行する（ステップＳ３１）。

【００７２】HDコントローラ１２０-i（ｉ＝1 ～ｎ）は、制御部１５から与えられる命令（read命令）で要求されたHDD１１０-ij （ｊは１～ｍのいずれか）からのデータ読み出しを行い（ステップＳ３２）、正常に読めたか否かをチェックする（ステップＳ３３）。

【００７３】もし、エラーがなかったならば、HDコントローラ１２０-iはステップＳ３２で読み出したデータをバッファ部１３内のバッファ１３０-iに格納し、制御部１５に正常終了を通知する（ステップＳ３４）。これに対してエラーが有ったならば、HDコントローラ１２０-iはリカバリを行わず、制御部１５にリードエラーを通知する（ステップＳ３５）。

【００７４】制御部１５は、各HDコントローラ１２０-iからのエラー通知を監視しており（ステップＳ３６）、エラーを通知したHDコントローラ１２０-iの数が０であるならば、バッファ部１３内のバッファ１３０-1～１３０-nには、HDコントローラ１２０-1～１２０-nに要求したデータが格納されているものと判断してステップＳ３７に進む。このステップＳ３７では、制御部１５はＲＡＩＤ機構１４に対してＲＡＩＤ３の機能を起動させ、バッファ１３０-1～１３０-(n-1)のデータを合成させた後、その合成後のデータをホスト装置に出力する。

【００７５】また、エラーを通知したHDコントローラ１２０-iの数が１であるならば、制御部１５はステップＳ３８に進む。このステップＳ３８では、制御部１５はＲＡＩＤ機構１４に対してＲＡＩＤ３の機能を起動させ、エラーを通知した唯一のHDコントローラ１２０-iを除くHDコントローラ１２０-1～１２０-nに対応するバッファ１３０-1～１３０-nのデータから、エラーを通知したHDコントローラ１２０-iが読み出すべきデータをバッファ１３０-i内にリカバリさせ、バッファ１３０-1～１３０-(n-1)のデータを合成させた後、その合成後のデータをホスト装置に出力する。

【００７６】また、エラーを通知したHDコントローラ１２０-iの数が２以上であるならば、制御部１５はＲＡＩＤ機構１４のＲＡＩＤ３機能を起動してもリカバリ不可能であるとして、その旨をホスト装置に通知してエラー終了する（ステップＳ３９）。

（４）ＲＡＩＤ３領域３１２への書き込み

次に、ＲＡＩＤ３領域３１２への書き込みについて、図７のフローチャートを参照して説明する。

【００７７】制御部１５は、ホスト装置からの指示が、監視システムで適用されるビデオデータ等、短期間保存される大量データのＲＡＩＤ３領域３１２への書き込みを行うものである場合、まずＲＡＩＤ機構１４に対してＲＡＩＤ３の機能を起動させる。これによりＲＡＩＤ機構１４は、ホスト装置からのデータ（書き込みデータ）

をＲＡＩＤ３でストライピングして、そのパリティを生成しながら、そのストライピングしたデータ及びパリティを、バッファ部１３内のそれぞれ対応するバッファ１３０-1～１３０-nに格納する（ステップＳ４１）。

【００７８】次に制御部１５は、ＨＤコントローラ部１２内の各ＨＤコントローラ１２０-1～１２０-nに対し、バッファ１３０-1～１３０-nに格納されたデータをＲＡＩＤ４でのパリティを生成せずに、ＲＡＩＤ４グループ２４-1～２４-n内の目的とするＨＤＤ１１０-1j ～１１０-nj （ｊは１～mのいずれか）に書き込むことを指示する（ステップＳ４２）。

【００７９】これを受けて各ＨＤコントローラ１２０-i（ｉ＝１～ｎ）は、バッファ１３０-iのデータをＲＡＩＤ４グループ２４-i内の目的とするＨＤＤ１１０-ij （ｊは１～mのいずれか）に書き込む（ステップＳ４３）。

（５）ＲＡＩＤ４領域３１３からの読み出し

次に、ＲＡＩＤ４領域３１３からの読み出しについて、図８のフローチャートを参照して説明する。

【００８０】制御部１５は、ホスト装置からの指示が、静止画等の小量データのＲＡＩＤ４領域３１３からの読み出しである場合、目的とするＨＤＤ１１０-ij （ｉは１～ｎのいずれか、ｊは１～mのいずれか）を含むＲＡＩＤ４グループ２４-iに対応するＨＤコントローラ部１２内のＨＤコントローラ１２０-iに対して、ＲＡＩＤ４によるリカバリ機能有りのリードのための命令（read命令）を発行する（ステップＳ５１）。

【００８１】ＨＤコントローラ１２０-iは、制御部１５から与えられる命令（read命令）で要求されたＨＤＤ１１０-ij からのデータ読み出しを行い（ステップＳ５２）、正常に読めたか否かをチェックする（ステップＳ５３）。

【００８２】もし、エラーがなかったならば、ＨＤコントローラ１２０-iはステップＳ５２で読み出したデータをバッファ部１３内のバッファ１３０-iに格納し、制御部１５に正常終了を通知する（ステップＳ５４）。

【００８３】これに対し、エラーが有ったならば、ＨＤコントローラ１２０-iはＲＡＩＤ４によるリカバリを実行する（ステップ５５）。即ちＨＤコントローラ１２０-iは、ＲＡＩＤ４グループ２４-i内のＨＤＤ１１０-ij を除く全てのＨＤＤ１１０-il ～１１０-im を対象に、ステップＳ５２で行ったのと同じ読み出しを行い、全て正常に読めたならば、リカバリ可能であるものと判断し、その読み出したデータの排他的論理和をとってＨＤＤ１１０-ij のデータをリカバリする。

【００８４】ＨＤコントローラ１２０-iは、ＲＡＩＤ４によるリカバリができた場合（ステップＳ５６）、リカバリしたデータをバッファ部１３内のバッファ１３０-iに格納し、制御部１５に正常終了を通知する（ステップＳ５７）。これに対し、正常に読めなかったＨＤＤが１つでも存在するならば、ＨＤコントローラ１２０-iはＲＡＩＤ４によるリカバリはできないものとして、制御部１５に対して異常終了（リードエラー）を通知する（ステップＳ５８）。

【００８５】制御部１５は、ＨＤコントローラ１２０-iからの終了通知を監視しており（ステップＳ５９）、正常終了が通知された場合には、ＲＡＩＤ機構１４に対してＲＡＩＤ３機能を起動させず、そのままバッファ１３０-iのデータをホスト装置に出力する（ステップＳ６０）。これに対し、異常終了が通知された場合には、制御部１５はリカバリ不可能あるとして、その旨をホスト装置に通知してエラー終了する（ステップＳ６１）。

（６）ＲＡＩＤ４領域３１３への書き込み

次に、ＲＡＩＤ４領域３１３への書き込みについて、図９のフローチャートを参照して説明する。

【００８６】制御部１５は、ホスト装置からの指示が、静止画等の小量データのＲＡＩＤ４領域３１３への書き込みである場合、まずホスト装置からのデータ（書き込みデータ）を、目的とするＨＤＤ１１０-ij （ｉは１～ｎのいずれか、ｊは１～mのいずれか）と接続されているＨＤコントローラ１２０-iに対応するバッファ１３０-iに格納する（ステップＳ７１）。

【００８７】次に制御部１５は、ＨＤコントローラ１２０-iに対し、バッファ１３０-iに格納されたデータを、ＲＡＩＤ４でのパリティを生成しながら、ＲＡＩＤ４グループ２４-i内の目的とするＨＤＤ１１０-ij に書き込むことを指示する（ステップＳ７２）。

【００８８】これを受けてＨＤコントローラ１２０-iは、ＲＡＩＤ４でのパリティを生成しながら、バッファ１３０-iのデータをＲＡＩＤ４グループ２４-i内の目的とするＨＤＤ１１０-ij に書き込むと共に、生成したパリティをＲＡＩＤ４グループ２４-i内のＨＤＤ１１０-im に書き込む（ステップＳ７３）。なお、このＲＡＩＤ４でのパリティの生成は、１つのＨＤコントローラ１２０-iだけで行われる点を除けば、前記した（２）のＲＡＩＤ３＆４領域３１１への書き込みにおけるステップＳ２３でのパリティの生成と同様である。

【００８９】さて、前記した図８のフローチャートに従うＲＡＩＤ４領域３１３からの読み出し（小量データの読み出し）では、目的とするＨＤＤ１１０-ij からのデータ読み出しが正常に行われなかったならば、そのリカバリのために、当該ＨＤＤ１１０-ij を含むＲＡＩＤ４グループ２４-i内の当該ＨＤＤ１１０-ij 以外の全てのＨＤＤ１１０-il ～１１０-im を対象とする読み出し命令が発生する。このとき、ホスト装置から制御部１５に対して新たなＩ／Ｏ要求が送られた場合、その要求は、上記のリカバリが終了するまで待たされる。しかし、新たなＩ／Ｏ要求の対象となるＨＤＤが、例えばＨＤＤ１１０-ij を含むＲＡＩＤ３グループ２３-jであるというように、並行してアクセス可能な場合でも、ＲＡＩＤ４

による小量データの読み出しのためにその要求が待たされるのは無駄である。

【００９０】そこで、このような不具合を解消するようにした、改良されたＲＡＩＤ４領域３１３からの読み出しにつき説明する。ここでは、制御部１５からＨＤコントローラ１２０-iに対し、ＲＡＩＤ４領域３１３からのデータ読み込み（小量データの読み込み）のためのリード命令を発行する際に、３つのモードの１つが指定可能なようになっている。この３つのモードは、（Ａ）ＲＡＩＤ４のリカバリ有りモード、（Ｂ）ＲＡＩＤ４のリカバリ無しモード、（Ｃ）ＲＡＩＤ４のリカバリモードである。以下、この３つのモードについて説明する。

（Ａ）ＲＡＩＤ４のリカバリ有りモード

このモードは、前記（５）のＲＡＩＤ４領域３１３からの読み出しと同様の動作を行うものであり、目的とするＨＤＤ１１０-ij （ｉは１～ｎのいずれか、ｊは１～ｍのいずれか）を含むＲＡＩＤ４グループ２４-i内の全てのＨＤＤ１１０-i1 ～１１０-im に対してリード命令が発生される可能性がある。

（Ｂ）ＲＡＩＤ４のリカバリ無しモード

このモードでは、目的とするＨＤＤ１１０-ij のみにリード命令が発生する。

（Ｃ）ＲＡＩＤ４のリカバリモード

このモードでは、目的とするＨＤＤ１１０-ij を含むＲＡＩＤ４グループ２４-i内のＨＤＤ１１０-i1 ～１１０-im のうち、目的とするＨＤＤ１１０-ij を除く全てのＨＤＤへのリード命令のみが発生する。ここでは、目的とするＨＤＤ１１０-ij でのエラーの有無に無関係に、当該ＨＤＤ１１０-ij を除くＲＡＩＤ４グループ２４-i内の全ＨＤＤから読み込んだデータにより、目的とするＨＤＤ１１０-ij のデータをリカバリする。したがって、このモードでは、目的とするＨＤＤ１１０-ij へのアクセスが発生しないことが保証される。但し、この（Ｃ）のモードでは読み込めなくて、上記（Ａ）のモードでは読み込めるケースがある。これについて、図１０を参照して説明する。

【００９１】まず、ＲＡＩＤ４グループ２４-i内のＨＤＤ１１０-i1 ～１１０-im のうち、図１０に示すようなＨＤＤ１１０-i1 内の領域１００-1からの読み込みがホスト装置から制御部１５に対して指示されたものとする。ここで領域（読み込み対象領域）１００-1は３セクタ分のサイズがあり、先頭セクタのみにエラー箇所があるものとする。また、ＨＤＤ１１０-i1 内の領域１００-1に対応するＨＤＤ１１０-i2 内の領域１００-2には、最終セクタのみにエラー箇所があり、ＨＤＤ１１０-i1 ～１１０-im のうち、ＨＤＤ１１０-i1 ，１１０-i2 を除く残りのＨＤＤの対応する領域には、エラー箇所は無いものとする。またＨＤＤからの読み込みは、最小の読み込み単位であるセクタ単位で行われるものとする。

【００９２】このような場合、上記（Ｃ）のリカバリモードでＨＤＤ１１０-i1 内の領域１００-1のデータを読み込もうとすると、領域１００-1の先頭セクタと２番目のセクタについては、ＨＤＤ１１０-i2 ～１１０-im の対応するセクタにエラー箇所が無いため、その対応するセクタの排他的論理和をとることで取得（リカバリ）可能である、これに対し、領域１００-1の最後のセクタについては、ＨＤＤ１１０-i2 の対応するセクタにエラー箇所が有るためエラーが発生し、失敗に終わる。勿論、ＨＤＤ１１０-i1 を除く残りのＨＤＤ１１０-i2 ～１１０-im の対応する領域にエラー箇所が無いならば、このＨＤＤ１１０-i2 ～１１０-im の対応する領域からの読み込みを行って、その排他的論理和をとることで、目的とするＨＤＤ１１０-i1 の領域１００-1内のデータに相当するデータを取得（リカバリ）することが可能である。

【００９３】一方、上記（Ａ）のリカバリ有りモードでＨＤＤ１１０-i1 の領域１００-1のデータを読み込む場合には、まず領域１００-1の先頭セクタを読み込もうとしてエラーが発生するが、ＨＤＤ１１０-i2 ～１１０-im の対応するセクタにエラー箇所が無いため、その対応するセクタの排他的論理和をとることでリカバリ可能である。領域１００-1の後続のセクタ（２番目と３番目のセクタ）については、エラー箇所が無く、ＨＤＤ１１０-i1 からの読み込みが正常に行われるため、正常に終了する。

【００９４】さて、本実施形態では、図１中の制御部１５内に、ＲＡＩＤ４領域３１３への入出力のための待ち行列であるキュー（キューエントリ列）を、各ＲＡＩＤグループ（ここでは、ＲＡＩＤ４グループ２４-1～２４-n）毎に有している。図１１に、１つのＲＡＩＤ４グループに対応するキュー（以下、キュー＃３と称する）の一例を示す。

【００９５】図１１において、ヘッダ部１１１に保持されたポインタ１１２で指定されるキューエントリ（先頭のキューエントリ）１１３-1は、リードであるか或いはライトであるかを示すフラグ（Ｒ／Ｗフラグ）１１３１、リード或いはライトの対象となるＨＤＤ、即ち目的とするＨＤＤ（ディスク装置）の識別子（装置識別子）１１３２、リード／ライトデータを一時的に格納するバッファ１３０-i内の格納先先頭番地１１３３、シークアドレス（装置識別子の示すＨＤＤ内のＲＡＩＤ４領域３１３のシークアドレス）１１３４、転送長１１３５、モード識別子１１３６、他のＩ／Ｏとの干渉が無い場合にかかる時間（想定される実行時間）の上限値１１３７、及び次のキューエントリを指す次エントリポインタ１１３８を有している。モード識別子１１３６は、Ｒ／Ｗフラグ１１３１がリードを示す場合には、ＲＡＩＤ４のリカバリ無しモード、ＲＡＩＤ４のリカバリモード、及びＲＡＩＤ４のリカバリ有りモードのいずれか１つを示す。また、モード識別子１１３６は、Ｒ／Ｗフラグ１１

３１がライトを示す場合には、常にパリティを生成するモードを示す。キューエントリ１１３-1につながる他のキューエントリ１１３-2…も、当該キューエントリ１１３-1と同様のデータ構造をなす。

【００９６】ここで、先頭のキューエントリ１１３-1に設定されているＩ／Ｏ要求の実行について、当該キューエントリ１１３-1が、ホスト装置から制御部１５に対してＲＡＩＤ４グループ２４-i中のＨＤＤ１１０-i1 の領域１００-1からのＲＡＩＤ４での読み込みが指示された結果生成された場合を例に、図１２のフローチャートを参照して説明する。このときキューエントリ１１３-1中のＲ／Ｗフラグ１１３１はリードを、装置識別子１１３２はＨＤＤ１１０-i1 を、シークアドレス１１３４はＨＤＤ１１０-i1 内の領域１００-1の先頭アドレスを、転送長１１３３は３セクタを、モード識別子１１３６はＲＡＩＤ４のリカバリ無しモードを示しているものとする。

【００９７】まず制御部１５は、キューエントリ１１３-1の内容に従って、ＲＡＩＤ４のリカバリ無しモードでの読み込みを行うようにＨＤコントローラ部１２内のＨＤコントローラ１２０-iを制御する（ステップＳ８１）。

【００９８】この場合、ＨＤコントローラ１２０-iはＨＤＤ１１０-i1 の指定領域１００-1からのデータ読み込みのみを行う。もし、読み込みエラーとなった場合には、制御部１５は実行中のキューエントリ１１３-1のモード識別子をＲＡＩＤ４のリカバリモードに更新すると共に、他のＩ／Ｏとの干渉が無い場合にかかる時間の上限値１１３７を当該ＲＡＩＤ４のリカバリモードに合わせて更新し、当該キューエントリ１１３-1をつなぎ直す再スケジューリングを行って、機会を待つ。この上限値1137の更新については後述する。

【００９９】制御部１５は、上記更新後のキューエントリ１１3-1が実行可能になると、当該エントリ１１３-1の内容に従って、ＲＡＩＤ４のリカバリモードでの読み込みを行うようにＨＤコントローラ部１２内のＨＤコントローラ１２０-iを制御する（ステップＳ８２）。

【０１００】この場合、ＨＤコントローラ１２０-iはＨＤＤ１１０-i1 の指定領域１００－１に対応する他のＨＤＤ１１０－ｉ２　～ＨＤＤ１１０-im の領域からのデータ読み込みを行い、正常に読み込めたなら、その排他的論理和をとることで、指定領域１００-1のデータに相当するデータを取得（リカバリ）する。一方、１つでも読み込みエラーが発生したならば、制御部１５は実行中のキューエントリ１１３-1のモード識別子をＲＡＩＤ４のリカバリ有りモードに更新すると共に、他のＩ／Ｏとの干渉が無い場合にかかる時間の上限値１１３７を当該ＲＡＩＤ４のリカバリ有りモードに合わせて更新し、当該キューエントリ１１３-1をつなぎ直して、機会を待つ。

【０１０１】ここで、上記ＲＡＩＤ４のリカバリ無しモードでの読み込みと、上記ＲＡＩＤ４のリカバリモードでの読み込みの期間、その読み込みの対象となっているＨＤＤを除くＨＤＤへのＩ／Ｏ要求は実行可能である。

【０１０２】制御部１５は、上記更新後のキューエントリ１１３-1が実行可能になると、当該エントリ１１３-1の内容に従って、ＲＡＩＤ４のリカバリ有りモードでの読み込みを行うようにＨＤコントローラ部１２内のＨＤコントローラ１２０-iを制御する（ステップＳ８３）。

【０１０３】この場合、ＨＤコントローラ１２０-iは、まずＨＤＤ１１０-i1 の指定領域１００-1からのデータ読み込みを行い、読み込みエラーが発生したならば、ＨＤＤ１１０-i1 を除く残りのＨＤＤ１１０-i2 ～１１０-im の対応する領域からの読み込みを行い、正常に読み込めたなら、その排他的論理和をとることで、目的とするＨＤＤ１１０-i1 の領域１００-1内のデータに相当するデータをリカバリする動作を、例えば読み込みの最小単位（ここではセクタ単位）で繰り返す。

【０１０４】このように本実施形態においては、ＲＡＩＤ４での読み込みを最初から（他のＩ／Ｏ要求が待たされる）リカバリ有りモードで行うのではなくて、まず（目的のＨＤＤ以外を対象とするＩ／Ｏ要求の並列実行が可能な）ＲＡＩＤ４のリカバリ無しモードで読み込みを行い、読み込みエラーが発生したならば、（目的のＨＤＤを対象とする別のＩ／Ｏ要求の並列実行が可能な）ＲＡＩＤ４のリカバリモードで読み込み行い、それでも読み込みエラーが発生した場合に、ＲＡＩＤ４のリカバリ有りモードでの読み込み行うようにしている。

【０１０５】即ち本実施形態においては、ＲＡＩＤ４での小量データの読み込みを、ビデオデータ等の大量データの読み込みのスケジューリングに極力影響を与えないような方法を優先的に適用して実行し、動作するＨＤＤを細かく制御しているため、大量データの読み込みのスケジューリングに影響を与えないようにしながら小量データの読み込みを行う機会を増加させることができる。

【０１０６】なお、制御部１５はスケジューリングの際、他のＩ／Ｏ要求、特にＲＡＩＤ３＆４領域３１１、ＲＡＩＤ３領域３１２への（大量データの）Ｉ／Ｏ要求の混み具合が少ない場合には、ステップＳ８１，Ｓ８２を省略する。

【０１０７】さて、先に述べた図４のフローチャートに従うＲＡＩＤ３＆４領域３１１からのリカバリ有りモードでの読み出しでは、各ＨＤコントローラ１２０-i（ｉ＝１～ｎ）は、制御部１５から与えられる命令（read命令）で要求されたＨＤＤ１１０-ij （ｊは１～mのいずれか）からのデータ読み出しを行い、正常に読めなかったなら、そのＨＤコントローラ１２０-iはＲＡＩＤ４グループ２４-i内のＨＤＤ１１０-i-1～１１０-im をアクセスしてＲＡＩＤ４でのリカバリ処理を行う。しかし、１台のＨＤコントローラ１２０-iのみがリカバリ処理を

行う場合には、バッファ１３０-1～１３０-mの内容をもとにＲＡＩＤ機構１４にてＲＡＩＤ３でのリカバリが可能なことから、その時間（ＲＡＩＤ４でのリカバリ処理の時間）が無駄となる。

【０１０８】そこで、このような不具合を解消するようにした、改良されたＲＡＩＤ３＆４領域３１１からのリカバリ有りモードでの読み出しにつき説明する。ここでは、前記したＲＡＩＤ４領域３１３への入出力を管理するためのキュー＃３とは別に、ＲＡＩＤ３＆４領域３１１及びＲＡＩＤ３領域３１２への入出力を管理するための、次に述べる２種類のキュー（キューエントリ列）を制御部１５内に有している。

【０１０９】図１３は、ＲＡＩＤ３＆４領域３１１及びＲＡＩＤ３領域３１２への入出力のための通常のＩ／Ｏ要求のキュー（以下、キュー＃１と称する）の一例を示す。このキュー＃１は、ＲＡＩＤ３のグループ（ここでは、ＲＡＩＤ３グループ２３-1～２３-m）の数だけ設けられる。

【０１１０】図１３において、ヘッダ部１３１に保持されたポインタ１３２で指定されるキューエントリ（先頭のキューエントリ）１３３-1は、リードであるか或いはライトであるかを示すフラグ（Ｒ／Ｗフラグ）１３３１、リード／ライトデータを一時的に格納するバッファ１３０-i内の格納先先頭番地１１３３、シークアドレス（ＲＡＩＤ３グループ２３-i内のシークアドレス）１１３４、転送長１１３５、モードフラグ１１３６、及び次のキューエントリを指す次エントリポインタ１１３８を有している。キューエントリ１３３-1につながる他のキューエントリ１３３-2…も、当該キューエントリ１３３-1と同様のデータ構造をなす。

【０１１１】モードフラグ１３３６は、Ｒ／Ｗフラグ１３３１がリードを示す場合には、元のＩ／Ｏ要求にＲＡＩＤ４のリカバリが必要か否かを示す。ここでは、ＲＡＩＤ３＆４領域３１１からのリカバリ有りモードでの読み出しの場合のみ、元のＩ／Ｏ要求にＲＡＩＤ４のリカバリが必要となる。また、モードフラグ１３３６は、Ｒ／Ｗフラグ１３３１がライトを示す場合にはＲＡＩＤ４のパリティ生成が必要か否かを示す。ここでは、ＲＡＩＤ３＆４領域３１１への書き込みの場合のみ、パリティ生成が必要となる。

【０１１２】図１４は、ＲＡＩＤ３＆４領域３１１に対するＲＡＩＤ４のリカバリ有りモードの読み込み、またはＲＡＩＤ４のパリティ生成（パリティ生成・書き込み）の要求のキュー（以下、キュー＃２と称する）の一例を示す。このキュー＃２は、ＲＡＩＤ４のグループ（ここでは、ＲＡＩＤ４グループ２４-1～２４-n）の数だけ設けられる。

【０１１３】図１４において、ヘッダ部１４１に保持されたポインタ１４２で指定されるキューエントリ（先頭のキューエントリ）１４３-1は、リード（ＲＡＩＤ３＆４領域３１１に対するＲＡＩＤ４のリカバリ有りモードの読み込み）であるか或いはライト（ＲＡＩＤ４のパリティ生成が必要な書き込み）であるかを示すフラグ（Ｒ／Ｗフラグ）１４３１、リード或いはライトの対象となるＨＤＤ、即ち目的とするＨＤＤ（ディスク装置）の識別子（装置識別子）１４３２、リード／ライトデータを一時的に格納するバッファ１３０-i内の格納先先頭番地１４３３、シークアドレス（装置識別子の示すＨＤＤ内のシークアドレス）１４３４、転送長１４３５、他のＩ／Ｏとの干渉が無い場合にかかる時間の上限値１４３７、及び次のキューエントリを指す次エントリポインタ１４３８を有している。キューエントリ１４３-1につながる他のキューエントリ１４３-2…も、当該キューエントリ１４３-1と同様のデータ構造をなす。

【０１１４】ここで、キュー＃１の先頭のキューエントリ１３３-1に設定されているＩ／Ｏ要求の実行について、当該キューエントリ１３３-1が、ホスト装置から制御部１５に対してＲＡＩＤ３＆４領域３１１からのリカバリ有りモードでの読み込みが指示された結果生成された場合を例に、図１５のフローチャートを参照して説明する。このときキューエントリ１３３-1中のＲ／Ｗフラグ１３３１はリードを、モードフラグ１３３６は元のＩ／Ｏ要求にＲＡＩＤ４のリカバリが必要であることを示しているものとする。

【０１１５】まず制御部１５は、キュー＃１の先頭エントリであるキューエントリ１３３-1の内容に従って、ＨＤコントローラ部１２内の各ＨＤコントローラ１２０-i（ｉ＝１～ｎ）に対してＲＡＩＤ４のリカバリ機能無しのリードのための命令（read命令）を発行し、結果を受け取る（ステップＳ９１）。

【０１１６】ＨＤコントローラ１２０-iは制御部１５からの読み込み命令で要求されたＨＤＤ１１０-ij （ｊは１～ｍのいずれか）からの読み出しのみを行い、正常に読み込めたなら、その読み込んだデータをバッファ１３０-iに格納し、制御部１５に正常終了を通知する。これに対し、ＨＤＤ１１０-ij から正常に読み込めなかったならば、ＨＤコントローラ１２０-iは制御部１５にエラーを通知する。

【０１１７】制御部１５は、上記ステップＳ９１で各ＨＤコントローラ１２０-iからの終了通知を受け取ると、エラーを通知したＨＤコントローラ１２０-iの数を調べる（ステップＳ９２）。

【０１１８】もし、エラーを通知したＨＤコントローラ１２０-iの数が０であるならば、制御部１５は、バッファ部１３内のバッファ１３０-1～１３０-nには、ＨＤコントローラ１２０-1～１２０-nに要求したデータが格納されているものと判断してステップＳ９３に進む。このステップＳ９３では、制御部１５はＲＡＩＤ機構１４に対してＲＡＩＤ３の機能を起動させ、バッファ１３０-1～１３０-(n-1)のデータを合成させた後、その合成後の

データをホスト装置に出力する。
【０１１９】また、エラーを通知したＨＤコントローラ１２０-iの数が１であるならば、制御部１５はステップＳ９４に進む。このステップＳ９４では、制御部１５はＲＡＩＤ機構１４に対してＲＡＩＤ３の機能を起動させ、エラーを通知した唯一のＨＤコントローラ１２０-iを除くＨＤコントローラ１２０-1～１２０-nに対応するバッファ１３０-1～１３０-nのデータから、エラーを通知したＨＤコントローラ１２０-iが読み出すべきデータをバッファ１３０-i内にリカバリさせ、バッファ１３０-1～１３０-(n-1)のデータを合成させた後、その合成後のデータをホスト装置に出力する。
【０１２０】また、エラーを通知したＨＤコントローラ１２０-iの数が２以上であるならば（このように、２つ以上のＨＤコントローラからエラーが通知される確率は、１つのＨＤコントローラがエラーを通知する確率よりもはるかに小さい）、制御部１５はキューエントリ１３３-1に対応する図１４に示したようなキューエントリ１４３-1を生成してキュー＃２につなぎ、そのキューエントリ１４３-1の内容に従って、エラーを通知した２つ以上のＨＤコントローラ１２０-iに対し、ＲＡＩＤ４のリカバリ機能有りのリードのための命令（read命令）を発行し、結果を受け取る（ステップＳ９５）。
【０１２１】制御部１５からＲＡＩＤ４のリカバリ機能有りのリードのための命令が与えられた２つ以上のＨＤコントローラ１２０-iは、制御部１５からの読み込み命令で要求されたＨＤＤ１１０-ij （ｊは１～ｍのいずれか）からの読み出しを行い、エラーが有ったなら、前記した図４中のステップＳ５と同様にＲＡＩＤ４によるリカバリを実行する。そしてＨＤコントローラ１２０-iは、リカバリできたなら、制御部１５に対して正常終了を通知し、リカバリできなかったなら、制御部１５に対してエラーを通知する。
【０１２２】制御部１５は、上記ステップＳ９５で各ＨＤコントローラ１２０-iからの終了通知を受け取ると、エラーを通知したＨＤコントローラ１２０-iの数を調べる（ステップＳ９６）。
【０１２３】もし、エラーを通知したＨＤコントローラ１２０-iの数が０であるならば、制御部１５は、バッファ部１３内のバッファ１３０-1～１３０-nには、ＨＤコントローラ１２０-1～１２０-nに要求したデータが格納されているものと判断して上記ステップＳ９３に進む。このステップＳ９３では、制御部１５はＲＡＩＤ機構１４に対してＲＡＩＤ３の機能を起動させ、バッファ１３０-1～１３０-(n-1)のデータを合成させた後、その合成後のデータをホスト装置に出力する。
【０１２４】また、エラーを通知したＨＤコントローラ１２０-iの数が１であるならば、制御部１５は上記ステップＳ９４に進む。このステップＳ９４では、制御部１５はＲＡＩＤ機構１４に対してＲＡＩＤ３の機能を起動させ、エラーを通知した唯一のＨＤコントローラ１２０-iを除くＨＤコントローラ１２０-1～１２０-nに対応するバッファ１３０-1～１３０-nのデータから、エラーを通知したＨＤコントローラ１２０-iが読み出すべきデータをバッファ１３０-i内にリカバリさせ、バッファ１３０-1～１３０-(n-1)のデータを合成させた後、その合成後のデータをホスト装置に出力する。
【０１２５】このように本実施形態においては、１つのＨＤコントローラ１２０-iだけがエラーを通知した場合には、エラーを通知した唯一のＨＤコントローラ１２０-iを除くＨＤコントローラ１２０-1～１２０-nに対応するバッファ１３０-1～１３０-nのデータから、エラーを通知したＨＤコントローラ１２０-iが読み出すべきデータをバッファ１３０-i内にリカバリできる。即ち、リトライ時間がかからずに目的データをバッファ１３０-iに読み出せる。しかも、２つ以上のＨＤコントローラからエラーが通知される確率は極めて小さいが、そのような場合でも、遅くはなるが、目的のデータを読み出すことが可能となる。したがって、システム全体としては、データ読み出し時間が短縮される。
【０１２６】さて、先に述べた図５のフローチャートに従うＲＡＩＤ３＆４領域３１１への書き込みでは、ＲＡＩＤ４でのパリティの生成のために、例えばＲＡＩＤ４グループ２４-iを例にとると、当該グループ２４-i内のＨＤＤ１１０-i1 ～１１０-i(m-1) の対応する領域からデータを読み出し、その排他的論理和をとってパリティを生成しなければならない。このため、ＲＡＩＤ３＆４領域３１１への書き込みに多大な時間を要し、例えばビデオカメラからのビデオ信号をディジタル化し、更にエンコードしたもの（ビデオエンコーダの出力）をリアルタイムでディスクアレイ装置のＲＡＩＤ３＆４領域３１１に格納することは困難である。
【０１２７】そこで、このような不具合を解消するようにした、改良されたＲＡＩＤ３＆４領域３１１への書き込みについて、図１６のフローチャートを参照して説明する。
【０１２８】まず制御部１５は、ホスト装置からＲＡＩＤ３＆４領域３１１への書き込みが指示された場合、前記キュー＃１につなぐべき（登録すべき）キューエントリを生成し、当該キュー＃１につなぐ（ステップＳ１０１）。ここでは、図１３に示したキューエントリ１３３-1が生成されてキュー＃１につながれたものとする。この場合、キューエントリ１３３-1のＲ／Ｗフラグ１３３１はライトを示し、モードフラグ１３３６は、ＲＡＩＤ４のパリティ生成が必要であることを示す。
【０１２９】制御部１５は、キューエントリ１３３-1に従ってＲＡＩＤ機構１４に対してＲＡＩＤ３の機能を起動させることでホスト装置からのデータとそのパリティをバッファ部１３内のそれぞれ対応するバッファ１３０-1～１３０-nに格納させると共に、各ＨＤコントローラ

１２０-l～１２０-mを制御することで、バッファ１３０-l～１３０-n内のデータ（ホスト装置からのデータまたはそのパリティ）を、指定のＲＡＩＤ３グループ２３-j（ｊは１～ｍ－１のいずれか）内のＨＤＤ１１０-lj ～ＨＤＤ１１０-nj に書き込むＲＡＩＤ３による書き込みを行わせる（ステップＳ１０２）。

【０１３０】制御部１５は、キューエントリ１３３-lに従うＲＡＩＤ３による書き込みが正常終了したならば、当該キューエントリ１３３-lをもとに前記したキュー＃２につなぐべきキューエントリを生成し、当該キュー＃２につなぐ（ステップＳ１０３）。この処理は、ＲＡＩＤ４グループ毎に行われる。ここでは、図１４に示したキューエントリ１４３-lが生成されてキュー＃２につながれたものとする。この場合、キューエントリ１４３-lのＲ／Ｗフラグ１３３１はライトを示す。

【０１３１】制御部１５は、生成したキューエントリ１４３-lをキュー＃２につなぐと、ＲＡＩＤ４でのパリティ生成と書き込みが行われなくとも、ホスト装置に正常終了を通知する（ステップＳ１０４）。

【０１３２】一方、ＲＡＩＤ４グループ毎のキュー＃２につながれた各キューエントリ１４３-lは後で取り出されて実行され、ＲＡＩＤ４でのパリティが生成される（ステップＳ１０５）。このＲＡＩＤ４グループ毎に生成されたパリティは、そのグループ内のＨＤＤ１１０-im （ｉ＝１～ｎ）に書き込まれる。

【０１３３】このように本実施形態では、ＲＡＩＤ３＆４領域３１１への書き込みを、ＲＡＩＤ３によるパリティ生成を行いながらのデータ及びパリティ書き込みと、ＲＡＩＤ４によるパリティ生成・書き込みとの２段階で行い、短時間で処理できる前者の書き込みが終了（正常終了）した段階で、ホスト装置に対して終了（正常終了）を通知する構成としている。このため、ＲＡＩＤ３＆４領域３１１への書き込み時の見かけ上の性能向上が可能となり、リアルタイムエンコードの出力をそのままディスクアレイ装置に格納するような場合にも、必要な性能が確保できる。

【０１３４】なお、本実施形態では、上記した後者の書き込み（ＲＡＩＤ４によるパリティ生成・書き込み）が終了するまでの間は、該当するデータはＲＡＩＤ３で保護されているだけであるが、これだけでもある程度の保証はでき、しかも一時的な状態であることから、問題はない。

【０１３５】但し、上記ステップＳ１０３では、生成したキューエントリ１４３-lを、同一キュー＃２内のどの「ＲＡＩＤ４のリカバリ有りモードの読み込み」を要求するキューエントリよりも前方につなぐ必要がある。その理由は、もし、ＲＡＩＤ４のリカバリ有りモードの読み込み→他のＲＡＩＤ３グループへの書き込み→ＲＡＩＤ４のリカバリ有りモードの読み込みのリトライ→パリティ生成・書き込みという順序になると、誤ったデータを読み込むことになるので、そのような順序になるのを防ぐためである。

【０１３６】また、ＲＡＩＤ４のリカバリ有りモードの読み込み→同じＲＡＩＤ３グループへの書き込み→ＲＡＩＤ４のリカバリ有りモードの読み込みのリトライ→パリティ生成・書き込みという順序になると、読み込み命令（read命令）発行時より後の時点のデータを読み込むことになるが、これを避けるのはホスト装置側の責任であるとする。

【０１３７】次に、図１のディスクアレイ装置全体のＩ／Ｏ要求に対するスケジューリングについて、図１７及び図１８のフローチャートを参照して説明する。まず本実施形態では、制御部１５内の不揮発性メモリ１５０の所定番地の領域に、前記第１の実施形態で述べたＢレジスタ１５１と同様に、次に述べる３つのレジスタ、即ちｔレジスタ１５２、Ｔレジスタ１５３及びＡレジスタ１５４が割り当てらる。

【０１３８】ｔレジスタ１５２は、制御部１５が使用する変数ｔを記憶するためのもので、Ｔレジスタ１５３の値（Ｔ）とＩ／Ｏ処理に実際に要した時間との差分を蓄積していくために使用される。

【０１３９】Ｔレジスタ１５３は、システム立ち上げ時にホスト装置からの制御命令で指定される値Ｔを記憶しておくためのものである。本実施形態では、キュー＃１への登録対象となるＩ／Ｏ要求は、その転送長が予め定められた一定の長さ以上の場合には、一定長を最大の転送長とする分割した要求として扱われ、複数のキューエントリが生成されてキュー＃１につながれるようになっている。その理由は、キュー＃１につながれる各エントリの内容（の示すＩ／Ｏ要求）を実行するのに想定される所要時間をほぼ一定とするためである。この値に予め定められた余裕時間を加えた時間が上記Ｔとしてホスト装置から与えられてＴレジスタ１５３に設定されることになる。

【０１４０】Ａレジスタ１５４は、システム立ち上げ時にホスト装置からの制御命令で指定される値Ａを記憶しておくためのものである。前記したように、キュー＃２，＃３では、キューエントリをつなぐ際に、必要なＩ／Ｏの回数、転送長等から所要時間の上限値を見積もり、そのエントリ中に、他のＩ／Ｏとの干渉が無い場合にかかる時間の上限値１１３７，１４３７として設定される。Ａレジスタ１５４に設定される値Ａには、キュー＃２，＃３に登録されるＩ／Ｏ要求の実行に要する時間の上限値（他のＩ／Ｏとの干渉が無い場合にかかる時間の上限値１１３７，１４３７）及びＴレジスタ１５３の内容（Ｔ）より大きい値が用いられる。

【０１４１】さて、図１のディスクアレイ装置では、システム立ち上げ時に、ｔレジスタ１５２に初期値０がセットされる（ステップＳ１１１）。やがてホスト装置から制御部１５に対して何らかのＩ／Ｏ要求が与えられた

ならば、制御部１５はそのＩ／Ｏ要求を取り込んで、そのＩ／Ｏ要求に適合したキュー用のキューエントリを生成し、対応するキュー（キュー＃１～＃３のいずれか）につなぐ（ステップＳ１１２）。

【０１４２】次に制御部１５は、ｔレジスタ１５２の値ｔとＡレジスタ１５４の値Ａの大小を比較し、ｔ＜Ａの条件が成立するか否かをチェックする（ステップＳ１１３）。この時点では、上記ステップＳ１１１の処理でｔ＝０となっているため、ｔ＜Ａの条件が成立する。

【０１４３】すると制御部１５は、各ＲＡＩＤ３グループ２３-1～２３-m毎に用意されるキュー＃１から、ＲＡＩＤ３のグループが重複しないように、キューエントリを取り出し、スケジュールする（ステップＳ１１４）。

【０１４４】次に制御部１５は、各ＲＡＩＤ４グループ２４-1～２４-n毎に用意されるキュー＃２から、所要時間の上限値（他のＩ／Ｏとの干渉が無い場合にかかる時間の上限値１４３７）がＴレジスタ１５３の値Ｔより小さく且つＲＡＩＤ４のグループが重複しないようなキューエントリを取り出し、スケジュールする（ステップＳ１１５）。

【０１４５】次に制御部１５は、各ＲＡＩＤ４グループ２４-1～２４-n毎に用意されるキュー＃３から、所要時間の上限値（他のＩ／Ｏとの干渉が無い場合にかかる時間の上限値１１３７）がＴレジスタ１５３の値Ｔより小さく且つ上記ステップＳ１１４，Ｓ１１５でスケジュールしたＩ／ＯとＨＤＤが重複しないようなキューエントリを取り出し、スケジュールする（ステップＳ１１６）。

【０１４６】次に制御部１５は、ステップＳ１１３～Ｓ1.16で少なくとも１つスケジュールしたか否かをチェックし（ステップＳ１１７）、少なくとも１つスケジュールしたときには、ｔレジスタ１５２の値ｔを、その値ｔにＴレジスタ１５３の値Ｔを加えたｔ＋Ｔ（ここではＴ）に更新する（ステップＳ１１８）。

【０１４７】次に制御部１５は、ステップＳ１１３～Ｓ１１６でスケジュールしたＩ／Ｏ要求を各ＨＤコントローラ１２０-iに発行し、全て終了するまで待つ（ステップＳ１１９）。そして制御部１５は、ステップＳ１１９で要求したＩ／Ｏが全て終了すると、ｔレジスタ１５２の値ｔを、その値ｔ（ここではｔ＝Ｔ）から当該Ｉ／Ｏの経過時間（所用時間）を差し引いた値に更新し（ステップＳ１２０）、ステップＳ１１２に戻る。ここで、Ｔ＞Ｉ／Ｏの経過時間であることから、上記ステップＳ１１８～Ｓ１２０の実行後のｔレジスタ１５２の値ｔは増加する。また、実際に経過する時間はＴ未満である。

【０１４８】以上により、ステップＳ１１８～Ｓ１２０を繰り返し実行すると、やがて、ステップＳ１１３の条件（ｔ＜Ａ）が成立しなくなる。また、ステップＳ１１３～Ｓ１１６で１つもスケジュールできなかった場合には、ｔレジスタ１５２の値ｔがＡレジスタ１５４の値Ａに更新されて（ステップＳ１２１）、ステップＳ１１２に戻るため、直ちにステップＳ１１３の条件（ｔ＜Ａ）が成立しなくなる。

【０１４９】このように本実施形態においては、ステップＳ１１２からステップＳ１２０までのループは、いずれ終了する。さて、ステップＳ１１３の条件（ｔ＜Ａ）が成立しなくなると、制御部１５はステップＳ１２２以降の処理に進む。即ち制御部１５は、まず全てのキュー＃２から、ＲＡＩＤ４のグループが重複しないように、キューエントリを取り出し、スケジュールする（ステップＳ１２２）。

【０１５０】次に制御部１５は、全てのキュー＃３から、上記ステップＳ１２２でスケジュールしたＩ／ＯとＨＤＤが重複しないようなキューエントリを取り出し、スケジュールする（ステップＳ１２３）。

【０１５１】次に制御部１５は、ステップＳ１２２，Ｓ１２３で少なくとも１つスケジュールしたか否かをチェックし（ステップＳ１２４）、少なくとも１つスケジュールしたときには、ステップＳ１２２，Ｓ１２３でスケジュールしたＩ／Ｏ要求を各ＨＤコントローラ１２０-iに発行し、全て終了するまで待つ（ステップＳ１２５）。そして制御部１５は、ステップＳ１２５で要求したＩ／Ｏが全て終了すると、ｔレジスタ１５２の値ｔを初期値０に更新し（ステップＳ１２６）、ステップＳ１１２に戻る。また、ステップＳ１２２，Ｓ１２３で１つもスケジュールできなかった場合には、制御部１５はステップＳ１２５をスキップしてステップＳ１２６に進み、ｔレジスタ１５２の値ｔを初期値０に更新する。

【０１５２】このように本実施形態においては、ステップＳ１２２以降の処理が実行されると、最終的にステップＳ１２６でｔレジスタ１５２の値ｔが０となり、上記ステップＳ１１３の条件（ｔ＜Ａ）が成立する。したがって、ステップＳ１２２以降の処理が続けて実行されることはない。このステップＳ１２２以降の処理を実行した場合、実際の経過時間は、Ａ未満である。

【０１５３】以上の説明から明らかなよう、本実施形態におけるスケジューリングでは、キュー＃１が空きの状態で新しいエントリが登録された場合には、そのエントリ（の示すＩ／Ｏ要求）は（Ａ＋Ｔ）時間以内にスケジュールされる。また１つのキュー＃１に例えばｐ個のエントリがつながっている状態で、新しいエントリが登録された場合には、先頭のエントリがスケジュールされるまでの最長時間は（Ａ＋Ｔ）であり、ｐ個のエントリのＩ／Ｏに要する時間の上限はｐＴであることから、新たに登録されたエントリ（の示すＩ／Ｏ要求）は（Ａ＋（ｐ＋１）Ｔ）時間以内にスケジュールされる。即ち、キュー＃１へのＩ／Ｏの頭出し時間が保証できる。

【０１５４】また、本実施形態におけるスケジューリングでは、キュー＃１が空にならないようにホスト装置から制御することより、Ｔ時間に１回の割合でＩ／Ｏ要求

を実行することができる。即ち、キュー＃１へのＩ／Ｏのレートが保証できる。但し、リカバリ有りモードのＩ／Ｏ要求を発行した場合で、キュー＃２につなぎかえられた場合には、その終了の通知は遅れる。

【０１５５】また、本実施形態におけるスケジューリングでは、キュー＃１にエントリがなくならない状態でも、キュー＃２，＃３のエントリにはスケジュールされる機会が訪れる。

【０１５６】このように本実施形態によれば、次のような効果を得ることができる。

（１）ＲＡＩＤ３＆４領域３１１及びＲＡＩＤ３領域３１２へのＩ／Ｏ（入出力）において、Ｉ／Ｏ要求の発行及び終了を監視し、キューの状態をホスト装置から想定することにより、今、発行しようとするＩ／Ｏ要求に対する最悪の実行開始時点を、ＲＡＩＤ４領域３１３への、（その後発行されるＩ／Ｏ要求も含めた）Ｉ／Ｏ状況に無関係に予測できる。

（２）ＲＡＩＤ３＆４領域３１１及びＲＡＩＤ３領域３１２へのＩ／Ｏにおいて、ホスト装置からＩ／Ｏ要求を十分に先出しすることにより、該当するＲＡＩＤ３グループのＨＤＤのＩ／Ｏのレートを保証することができる。但し、ＲＡＩＤ３＆４領域３１１からの読み込みの場合に、ＲＡＩＤ４のリカバリ有りモードのＩ／Ｏ要求を発行した場合、その終了通知は遅れることがあり得る。また、各Ｉ／Ｏの終了は一定間隔ではなく、前倒しで終了することがあり得る。

（３）ＲＡＩＤ３＆４領域３１１及びＲＡＩＤ３領域３１２へのＩ／Ｏのための要求が頻繁に発行される状況でも、ＲＡＩＤ４領域３１３へのＩ／Ｏは必ず「暇をみて」行われる。この場合、（要求するデータ転送長の上限を小さくするようにするなどして）Ｉ／Ｏの所要時間の上限値を極力減少させるように、ホスト装置から制御する必要が生じるが、ホスト装置から制御コマンドを発行することで、（Ａの値を小さくし）この頻度を増加させることも可能である。但し、上記（１）の予測時間を減少させる効果を生むので、ホスト装置側では、慎重にＡの値を設定する必要がある。

【０１５７】以上に述べたように、本実施形態によれば、ビデオデータのような大量のデータの入出力については、一定の入出力レートを保証し、且つ余裕があれば、静止画などの小量データの入出力も並行して実行されるディスクアレイ装置を実現できる。

【０１５８】以上に述べた実施形態では、図１中のＨＤＤ群１１０におけるｙ方向（縦方向）の配列であるＨＤＤ１１０-11 ～１１０-n1，…，１１０-1m ～１１０-nm は、図２に示したように、それぞれＲＡＩＤ３グループ２３-1，…，２３-mを構成しているものとして説明したが、これに限るものではない。例えば、上記したｙ方向（縦方向）の配列であるＨＤＤ１１０-11 ～１１０-n1，…，１１０-1m ～１１０-nm も、図１９に示すように、ＲＡＩＤ４グループ２５-1，…，２５-mとするようにしても構わない。この構成では、図１中の制御部１５に、ＲＡＩＤ３の実行機能に代えてＲＡＩＤ４の実行機能を持たせる必要がある点を除けば、前記第１の実施形態と同様である。

【０１５９】なお、図１９の構成では、１つのＨＤコントローラ１２０-i（ｉ＝１～ｎ）下に１つＲＡＩＤ４グループ２４-iが接続されているが、全てのＨＤコントローラ１２０-i（ｉ＝１～ｎ）に同数の複数のグループが接続される構成であっても構わない。

【０１６０】図１９のような構成とした場合、Ｉ／Ｏエラーのリカバリが必要になったとき、ＲＡＩＤ機構１４によるバッファ部１３のアクセス頻度が増大するものの、（この構成ではＲＡＩＤ４＆４領域と呼ぶべき）ＲＡＩＤ３＆４領域３１１に静止画等の小量データを配置することが可能となり、小量データに２重のＲＡＩＤ保護を行うことが可能となる。

【０１６１】

【発明の効果】以上詳述したように本発明によれば、同一ディスク装置上で複数のディスクアレイ方式を可能とすると共に、２重のディスクアレイ保護を可能とし、信頼性の向上を図ると共に、異種データの混在を可能とし、しかも各データ種類・用途に応じた最適なアクセスが行える。

【図面の簡単な説明】

【図１】本発明の一実施形態に係るマルチメディアサーバ用ディスクアレイ装置の構成を示すブロック図。

【図２】図１中のＨＤＤ群１１におけるＲＡＩＤ３及びＲＡＩＤ４のグループの構成例を示す図。

【図３】図１中の各ＨＤＤ１１０-ij（ｉ＝１～ｎ，ｊ＝１～ｍ）に装着されるディスクの領域分割例を示す図。

【図４】同実施形態におけるＲＡＩＤ３＆４領域からのリカバリ有りモードでの読み出しを説明するためのフローチャート。

【図５】同実施形態におけるＲＡＩＤ３＆４領域への書き込みを説明するためのフローチャート。

【図６】同実施形態におけるＲＡＩＤ３領域からの読み出しと、ＲＡＩＤ３＆４領域からのリカバリ無しモードでの読み出しを説明するためのフローチャート。

【図７】同実施形態におけるＲＡＩＤ３領域への書き込みを説明するためのフローチャート。

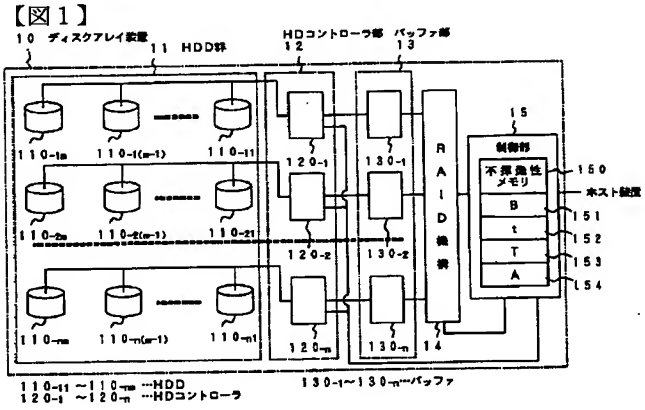
【図８】同実施形態におけるＲＡＩＤ４領域からの読み出しを説明するためのフローチャート。

【図９】同実施形態におけるＲＡＩＤ４領域への書き込みを説明するためのフローチャート。

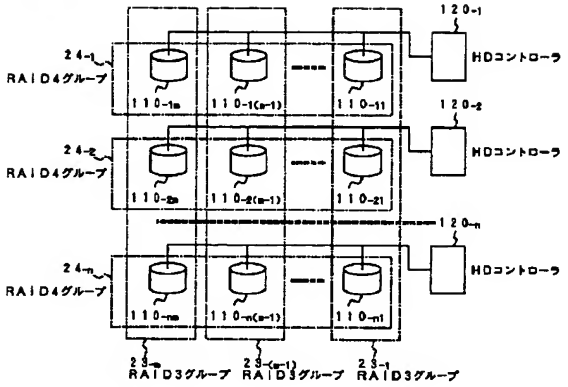
【図１０】図８に示したＲＡＩＤ４領域からの読み出しを改良するために用意される３つのモードのうちの、ＲＡＩＤ４のリカバリ有りモードとＲＡＩＤ４のリカバリモードとの違いを説明するための図。

【図１１】上記３つのモードを適用する場合のＲＡＩＤ４領域への入出力のためのキュー（＃３）の構造例を示す図。
【図１２】図８に示したＲＡＩＤ４領域からの読み出しを改良したフローチャート。
【図１３】図４に示したＲＡＩＤ３＆４領域からのリカバリ有りモードでの読み出しを改良するために適用される、ＲＡＩＤ３＆４領域及びＲＡＩＤ３領域への入出力を管理するためのキュー（＃１）の構造例を示す図。
【図１４】図４に示したＲＡＩＤ３＆４領域からのリカバリ有りモードでの読み出しを改良するために適用される、当該読み出し、またはＲＡＩＤ４のパリティ生成を要求するためのキュー（＃２）の構造例を示す図。
【図１５】図４に示したＲＡＩＤ３＆４領域からのリカバリ有りモードでの読み出しを改良したフローチャート。
【図１６】図５に示したＲＡＩＤ３＆４領域への書き込みを改良したフローチャート。
【図１７】同実施形態におけるＩ／Ｏ要求に対するスケジューリングを説明するためのフローチャートの一部を示す図。
【図１８】同実施形態におけるＩ／Ｏ要求に対するスケジューリングを説明するためのフローチャートの残りを示す図。
【図１９】図２の構成の変形例を示す図。
【図２０】従来のディスクアレイ装置のブロック構成図。
【図２１】従来のディスクアレイ装置のブロック構成図。
【符号の説明】
１０…ディスクアレイ装置、
１１…ＨＤＤ群（ディスク装置の群）、
１２…ＨＤコントローラ部、
１３…バッファ部、
１４…ＲＡＩＤ機構（入出力手段）、
１５…制御部、
２３-1～２３-m…ＲＡＩＤ３グループ（第１のディスクアレイグループ）、
２４-1～２４-n，２５-1～２５-m…ＲＡＩＤ４グループ（第２のディスクアレイグループ）、
１１０-11 ～１１０-nm，１１０-ij …ＨＤＤ（ディスク装置）、
１２０-1～１２０-n…ＨＤコントローラ（ディスクコントローラ）、
１３０-1～１３０-n…バッファ、
１５０…不揮発性メモリ、
１５１…Ｂレジスタ、
１５２…ｔレジスタ（第３の時間記憶手段）、
１５３…Ｔレジスタ（第１の時間記憶手段）、
１５４…Ａレジスタ（第２の時間記憶手段）、
３１１…ＲＡＩＤ３＆４領域（第１の領域）、
３１２…ＲＡＩＤ３領域（第２の領域）、
３１３…ＲＡＩＤ４領域（第３の領域）。

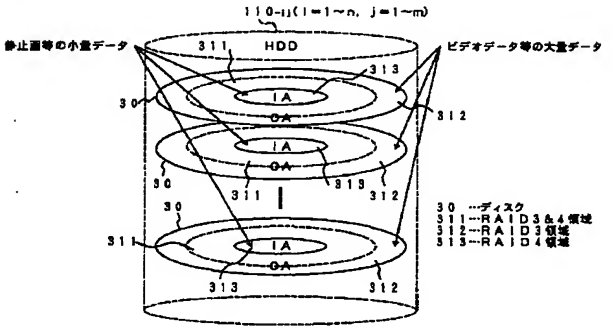
【図１】



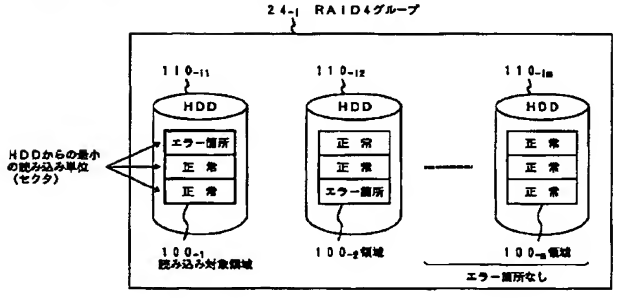
【図２】


120-1
HDコントローラ
24-1
RAID4グループ
110-1m
110-1(n-1)
110-11
120-2
HDコントローラ
24-2
RAID4グループ
110-2m
110-2(n-1)
110-21
120-m
HDコントローラ
24-m
RAID4グループ
110-mm
110-m(n-1)
110-m1
23-m
RAID3グループ
23-(n-1)
RAID3グループ
23-1
RAID3グループ

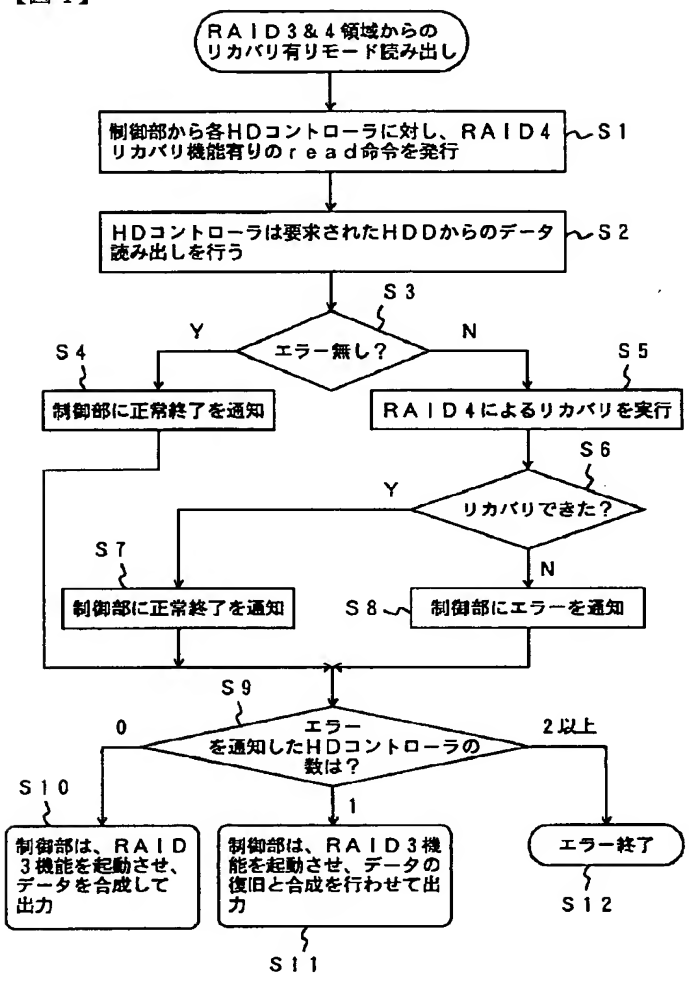
【図３】


110-ij(i=1～n, j=1～m)
静止画等の小容量データ
311
HDD
ビデオデータ等の大容量データ
313
30
IA
OA
312
30
311
313
312
30
311
313
312
30 …ディスク
311…RAID3&4領域
312…RAID3領域
313…RAID4領域

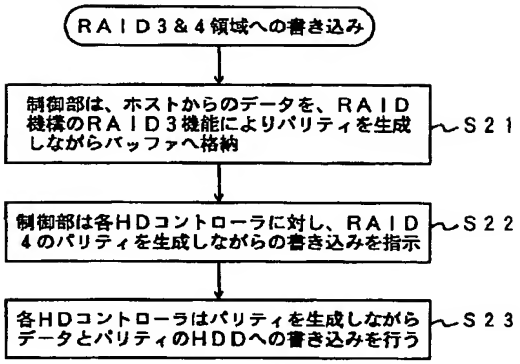
【図１０】


24-i RAID4グループ
110-i1
110-i2
110-im
HDD
HDD
HDD
エラー箇所
正常
正常
HDDからの最小の読み込み単位（セクタ）
正常
正常
正常
エラー箇所
正常
正常
正常
100-1
読み込み対象領域
100-2領域
100-m領域
エラー箇所なし

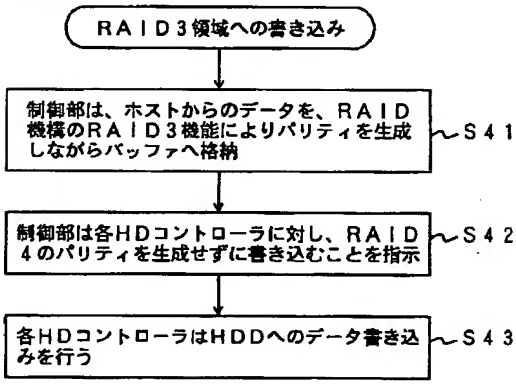
【図４】


ＲＡＩＤ３＆４領域からの
リカバリ有りモード読み出し
制御部から各ＨＤコントローラに対し、ＲＡＩＤ４
リカバリ機能有りのｒｅａｄ命令を発行
S１
ＨＤコントローラは要求されたＨＤＤからのデータ
読み出しを行う
S２
S３
エラー無し？
Y
N
S４
S５
制御部に正常終了を通知
ＲＡＩＤ４によるリカバリを実行
S６
リカバリできた？
Y
N
S７
制御部に正常終了を通知
S８
制御部にエラーを通知
S９
エラー
を通知したＨＤコントローラの
数は？
0
1
2以上
S１０
制御部は、ＲＡＩＤ
３機能を起動させ、
データを合成して
出力
制御部は、ＲＡＩＤ３機
能を起動させ、データの
復旧と合成を行わせて出
力
エラー終了
S１１
S１２

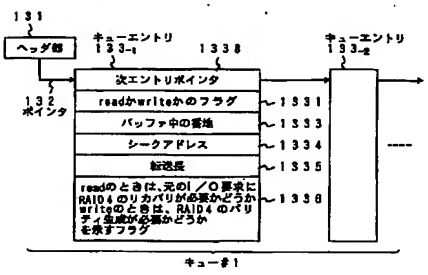
【図５】


ＲＡＩＤ３＆４領域への書き込み
制御部は、ホストからのデータを、ＲＡＩＤ機構のＲＡＩＤ３機能によりパリティを生成しながらバッファへ格納
Ｓ２１
制御部は各ＨＤコントローラに対し、ＲＡＩＤ４のパリティを生成しながらの書き込みを指示
Ｓ２２
各ＨＤコントローラはパリティを生成しながらデータとパリティのＨＤＤへの書き込みを行う
Ｓ２３

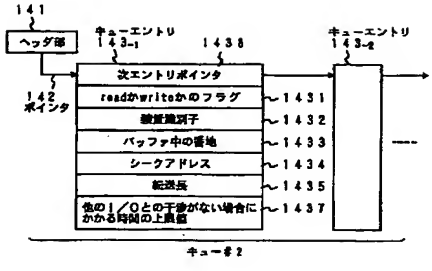
【図７】


ＲＡＩＤ３領域への書き込み
制御部は、ホストからのデータを、ＲＡＩＤ機構のＲＡＩＤ３機能によりパリティを生成しながらバッファへ格納
Ｓ４１
制御部は各ＨＤコントローラに対し、ＲＡＩＤ４のパリティを生成せずに書き込むことを指示
Ｓ４２
各ＨＤコントローラはＨＤＤへのデータ書き込みを行う
Ｓ４３

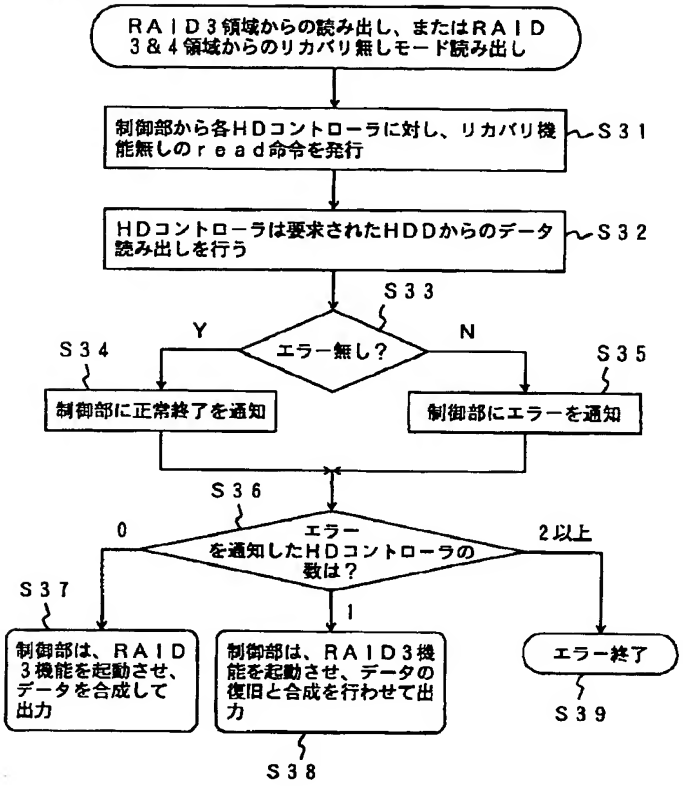
【図１３】


131
ヘッダ部
キューエントリ 133-1
1338
キューエントリ 133-2
132 ポインタ
次エントリポインタ
readかwriteかのフラグ
1331
バッファ中の番地
1333
シークアドレス
1334
転送長
1335
readのときは、元のＩ／Ｏ要求にRAID4のリカバリが必要かどうか writeのときは、RAID4のパリティ生成が必要かどうかを示すフラグ
1336
キュー#1

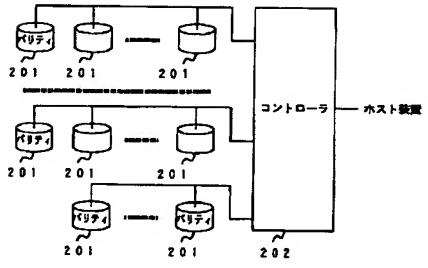
【図１４】


141
ヘッダ部
キューエントリ 143-1
1438
キューエントリ 143-2
142 ポインタ
次エントリポインタ
readかwriteかのフラグ
1431
装置識別子
1432
バッファ中の番地
1433
シークアドレス
1434
転送長
1435
他のＩ／Ｏとの干渉がない場合にかかる時間の上限値
1437
キュー#2

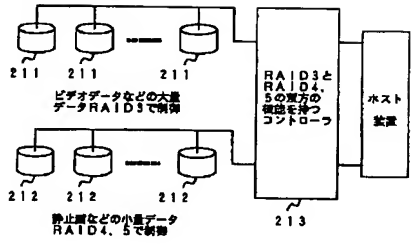
【図６】


ＲＡＩＤ３領域からの読み出し、またはＲＡＩＤ３＆４領域からのリカバリ無しモード読み出し
制御部から各ＨＤコントローラに対し、リカバリ機能無しのｒｅａｄ命令を発行
Ｓ３１
ＨＤコントローラは要求されたＨＤＤからのデータ読み出しを行う
Ｓ３２
Ｓ３３
エラー無し？
Ｙ
Ｎ
Ｓ３４
Ｓ３５
制御部に正常終了を通知
制御部にエラーを通知
Ｓ３６
エラーを通知したＨＤコントローラの数は？
０
１
２以上
Ｓ３７
制御部は、ＲＡＩＤ３機能を起動させ、データを合成して出力
制御部は、ＲＡＩＤ３機能を起動させ、データの復旧と合成を行わせて出力
エラー終了
Ｓ３８
Ｓ３９

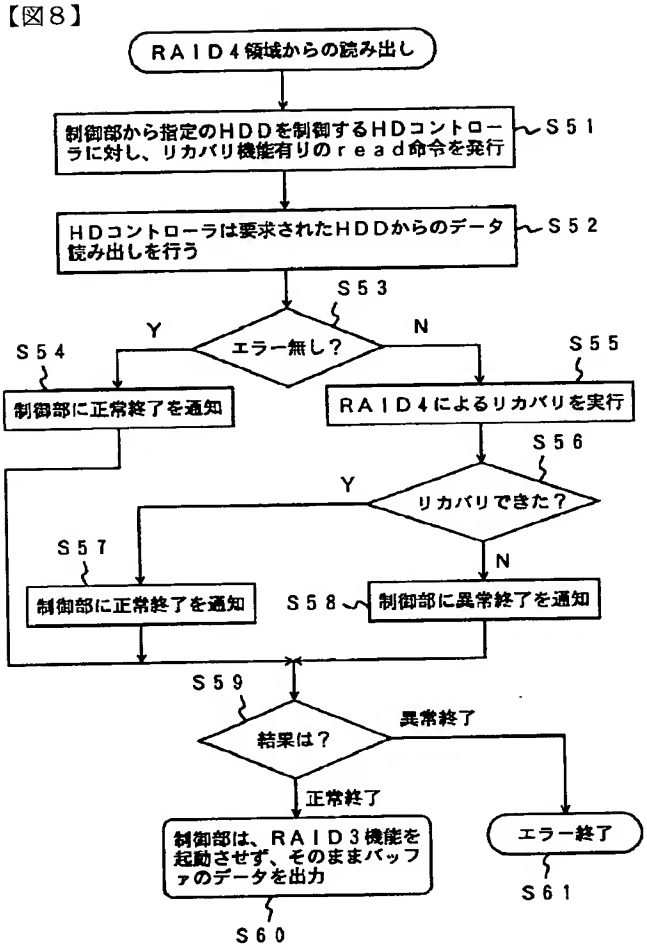
【図２０】


パリティ
201
202
コントローラ
ホスト装置

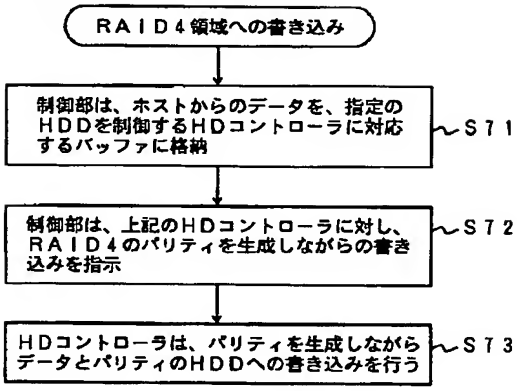
【図２１】


211
212
213
ビデオデータなどの大量データRAID3で制御
RAID3とRAID4、5の両方の機能を持つコントローラ
ホスト装置
静止画などの小量データRAID4、5で制御

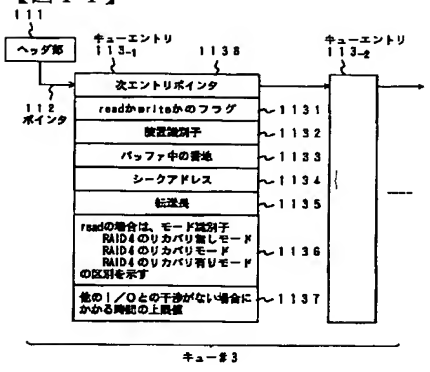
【図８】


ＲＡＩＤ４領域からの読み出し
制御部から指定のＨＤＤを制御するＨＤコントローラに対し、リカバリ機能有りのｒｅａｄ命令を発行
Ｓ５１
ＨＤコントローラは要求されたＨＤＤからのデータ読み出しを行う
Ｓ５２
Ｓ５３
エラー無し？
Ｙ
Ｎ
Ｓ５４
制御部に正常終了を通知
Ｓ５５
ＲＡＩＤ４によるリカバリを実行
Ｓ５６
リカバリできた？
Ｙ
Ｎ
Ｓ５７
制御部に正常終了を通知
Ｓ５８
制御部に異常終了を通知
Ｓ５９
結果は？
異常終了
正常終了
制御部は、ＲＡＩＤ３機能を起動させず、そのままバッファのデータを出力
Ｓ６０
エラー終了
Ｓ６１

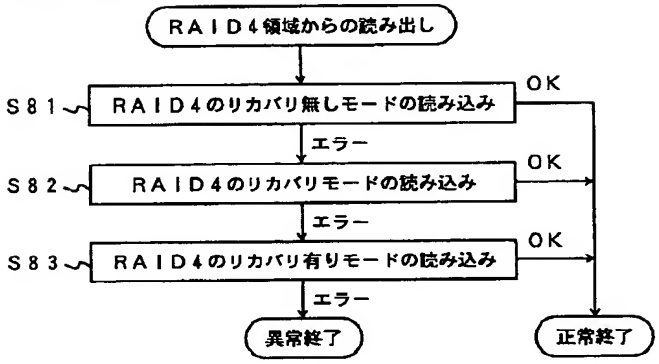
【図９】


RAID4領域への書き込み
制御部は、ホストからのデータを、指定のHDDを制御するHDコントローラに対応するバッファに格納
S71
制御部は、上記のHDコントローラに対し、RAID4のパリティを生成しながらの書き込みを指示
S72
HDコントローラは、パリティを生成しながらデータとパリティのHDDへの書き込みを行う
S73

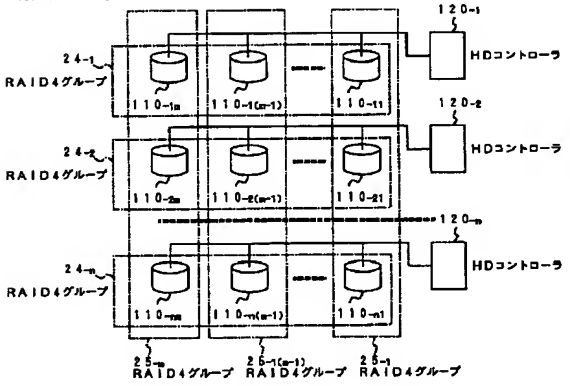
【図１１】


111
ヘッダ部
112
ポインタ
キューエントリ
113-1
1138
キューエントリ
113-2
次エントリポインタ
readかwriteかのフラグ
1131
1132
バッファ中の番地
1133
シークアドレス
1134
転送長
1135
readの場合は、モード識別子
RAID4のリカバリ無しモード
RAID4のリカバリモード
RAID4のリカバリ有りモード
の区別を示す
1136
他のI／Oとの干渉がない場合にかかる時間の上限値
1137
キュー#3

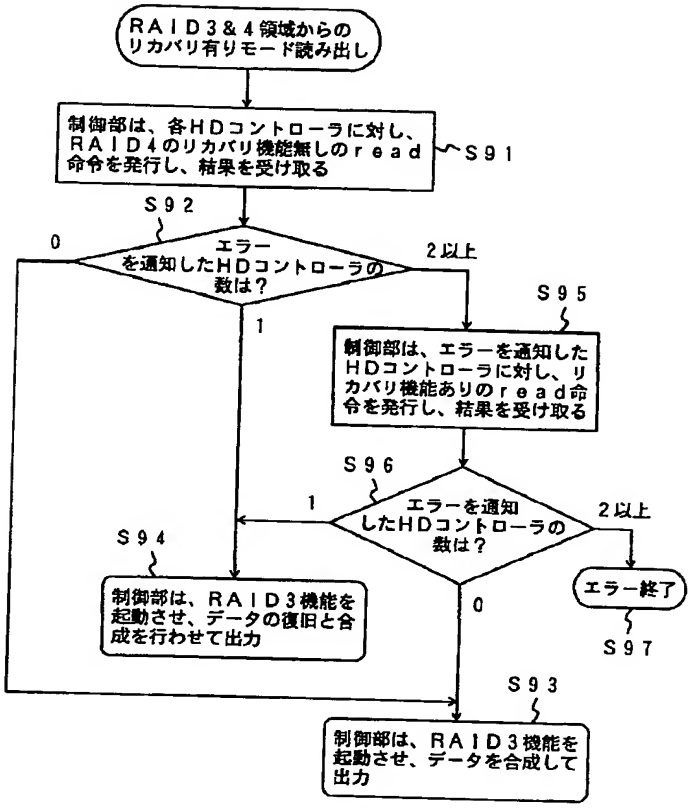
【図１２】


RAID4領域からの読み出し
S81
RAID4のリカバリ無しモードの読み込み
OK
エラー
S82
RAID4のリカバリモードの読み込み
OK
エラー
S83
RAID4のリカバリ有りモードの読み込み
OK
エラー
異常終了
正常終了

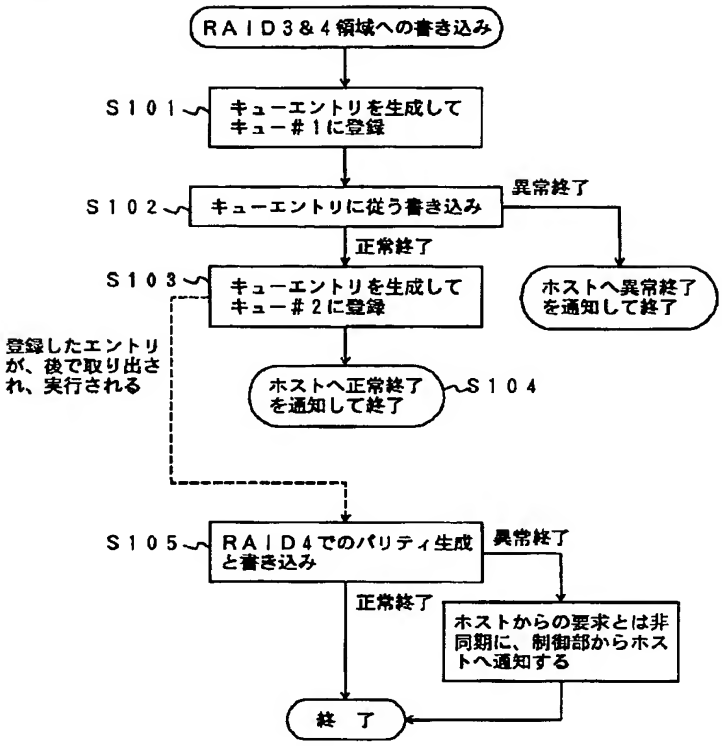
【図１９】


120-1
HDコントローラ
24-1
RAID4グループ
110-1m
110-1(m-1)
110-11
120-2
HDコントローラ
24-2
RAID4グループ
110-2m
110-2(m-1)
110-21
120-n
HDコントローラ
24-n
RAID4グループ
110-nm
110-n(m-1)
110-n1
25-m
RAID4グループ
25-1(m-1)
RAID4グループ
25-1
RAID4グループ

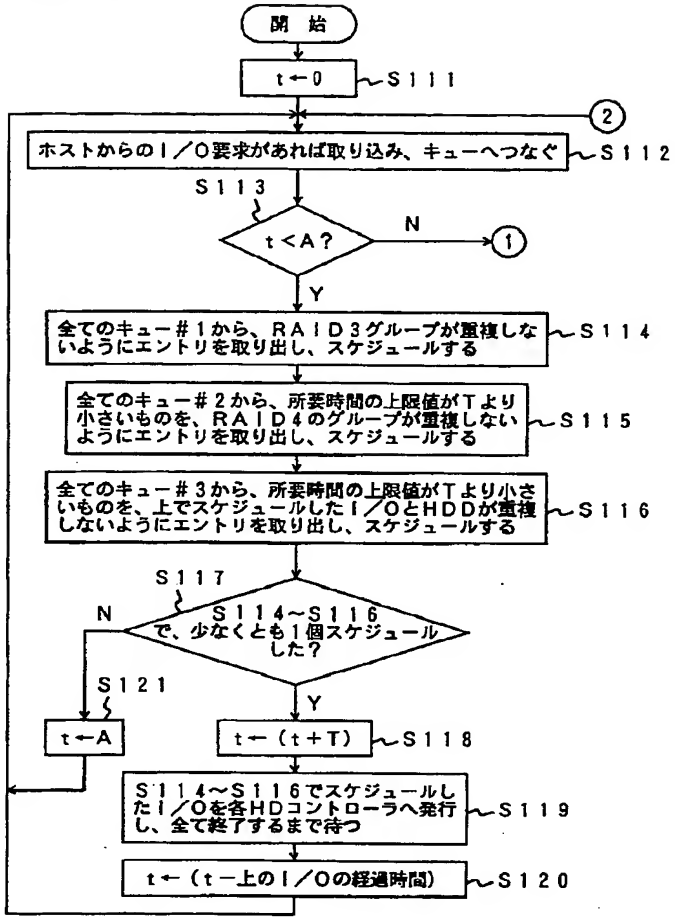
【図１５】


ＲＡＩＤ３＆４領域からのリカバリ有りモード読み出し
制御部は、各ＨＤコントローラに対し、ＲＡＩＤ４のリカバリ機能無しのｒｅａｄ命令を発行し、結果を受け取る
Ｓ９１
Ｓ９２
エラーを通知したＨＤコントローラの数は？
0
1
2以上
Ｓ９５
制御部は、エラーを通知したＨＤコントローラに対し、リカバリ機能ありのｒｅａｄ命令を発行し、結果を受け取る
Ｓ９６
エラーを通知したＨＤコントローラの数は？
1
2以上
0
Ｓ９４
制御部は、ＲＡＩＤ３機能を起動させ、データの復旧と合成を行わせて出力
エラー終了
Ｓ９７
Ｓ９３
制御部は、ＲＡＩＤ３機能を起動させ、データを合成して出力

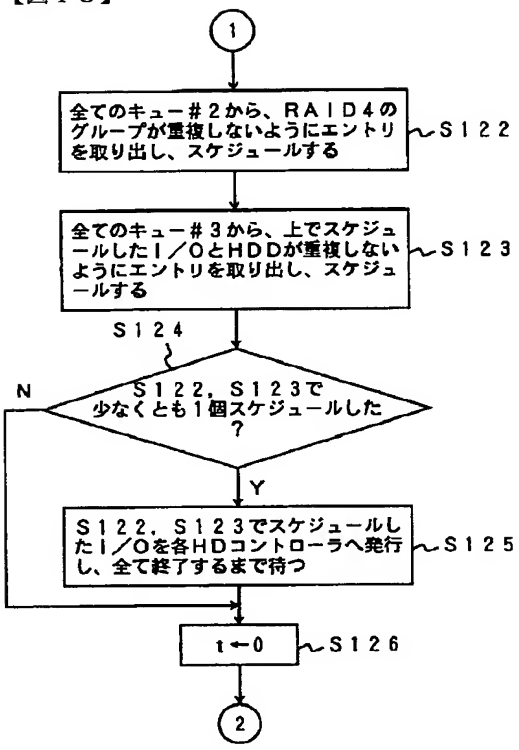
【図１６】


RAID3&4領域への書き込み
S101
キューエントリを生成して
キュー#1に登録
S102
キューエントリに従う書き込み
異常終了
正常終了
ホストへ異常終了
を通知して終了
S103
キューエントリを生成して
キュー#2に登録
登録したエントリ
が、後で取り出さ
れ、実行される
ホストへ正常終了
を通知して終了
S104
S105
RAID4でのパリティ生成
と書き込み
異常終了
正常終了
ホストからの要求とは非
同期に、制御部からホス
トへ通知する
終　了

【図１７】


開 始
t←0
S１１１
②
ホストからのＩ／Ｏ要求があれば取り込み、キューへつなぐ
S１１２
S１１３
t＜A？
N
①
Y
全てのキュー＃１から、ＲＡＩＤ３グループが重複しないようにエントリを取り出し、スケジュールする
S１１４
全てのキュー＃２から、所要時間の上限値がＴより小さいものを、ＲＡＩＤ４のグループが重複しないようにエントリを取り出し、スケジュールする
S１１５
全てのキュー＃３から、所要時間の上限値がＴより小さいものを、上でスケジュールしたＩ／ＯとＨＤＤが重複しないようにエントリを取り出し、スケジュールする
S１１６
S１１７
S１１４～S１１６で、少なくとも１個スケジュールした？
N
Y
S１２１
t←A
t←（t＋T）
S１１８
S１１４～S１１６でスケジュールしたＩ／Ｏを各ＨＤコントローラへ発行し、全て終了するまで待つ
S１１９
t←（t－上のＩ／Ｏの経過時間）
S１２０

【図１８】


1
全てのキュー＃２から、ＲＡＩＤ４のグループが重複しないようにエントリを取り出し、スケジュールする
Ｓ１２２
全てのキュー＃３から、上でスケジュールしたＩ／ＯとＨＤＤが重複しないようにエントリを取り出し、スケジュールする
Ｓ１２３
Ｓ１２４
Ｓ１２２，Ｓ１２３で少なくとも１個スケジュールした？
N
Y
Ｓ１２２，Ｓ１２３でスケジュールしたＩ／Ｏを各ＨＤコントローラへ発行し、全て終了するまで待つ
Ｓ１２５
t←0
Ｓ１２６
2